DOCUMENT RESUME

ED 090 948 IR 000 558

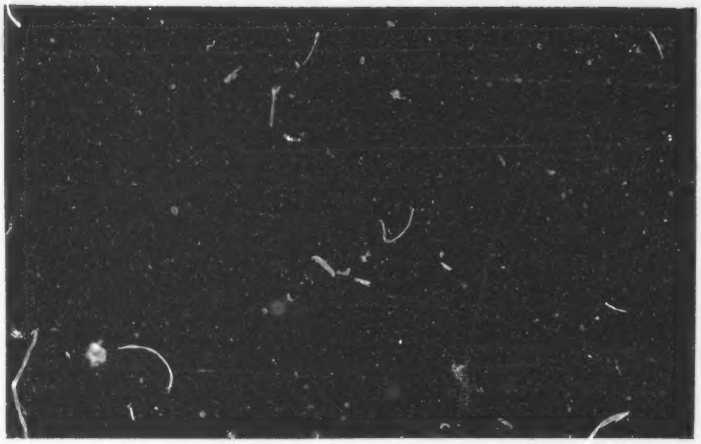
AUTHOR	Rush, J. E.; And Others
TITLE	A Computer Assisted Language Analysis System.
INSTITUTION	Ohio State Univ., Columbus. Computer and Information Science Research Center.
SPONS AGENCY	National Science Foundation, Washington, D.C.
REPORT NO	OSU-CISRC-TR-74-1
PUB DATE	Feb 74
NOTE	75p.
EDRS PRICE	MF-$0.75 HC-$4.20 PLUS POSTAGE
DESCRIPTORS	Case (Grammar); Computational Linguistics; *Computers; Interaction; *Language Patterns; Language Research; *Linguistic Patterns; Linguistics; *Oral Communication; Program Descriptions; Speech
IDENTIFIERS	CALAS; *Computer Assisted Language Analysis System; Dialogue Analysis; Natural Language

ABSTRACT

A description is presented of a computer-assisted language analysis system (CALAS) which can serve as a method for isolating and displaying language utterances found in conversation. The purpose of CALAS is stated as being to deal with the question of whether it is possible to detect, isolate, and display information indicative of what is happening in a conversation and ultimately to determine and predict outcomes of human interaction on the basis of speech patterns. Section I contains an introduction to CALAS and lists the criteria used in developing the system; the following section discusses conceptualizations of natural language interaction. Section III describes the method adopted for the representation of language presented in interaction as a modified form of case grammar. Section IV treats the mechanics of automated processing of natural language, while section V discusses CALAS as a basis for interpreting natural language. A brief final chapter sketches some future plans for the refinement of CALAS and its empirical use in dialogue analysis. (PB)

ED 090948

TECHNICAL REPORT SERIES



COMPUTER & INFORMATION SCIENCE RESEARCH CENTER

THE OHIO STATE UNIVERSITY COLUMBUS, OHIO

ED 090948

(OSU-CISRC-TR-74-1)

A COMPUTER ASSISTED LANGUAGE ANALYSIS SYSTEM

by

J. E. Rush, H. B. Pepinsky,
B. C. Landry, N. M. Meara,
S. M. Strong, J. A. Valley and C. E. Young

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Work performed under

Grant No. 534.1, National Science Foundation

Computer and Information Science Research Center
The Ohio State University
Columbus, Ohio 43210
February 1974

PREFACE

This work was supported in part by Grant No. GN 534.1 from the Office of Science Information Service, National Science Foundation to the Computer and Information Science Research Center of The Ohio State University. It was also supported by The Mershon Center, Programs of Research and Education in Leadership and Public Policy.

The Computer and Information Science Research Center of The Ohio State University is an interdisciplinary research organization which consists of the staff, graduate students and faculty of many University departments and laboratories. This report is based on research accomplished in cooperation with the Department of Computer and Information Science.

The research was administered and monitored by The Ohio State University Research Foundation.

TABLE OF CONTENTS

	Page
Preface	ii
List of Figures	v
List of Tables	v
1. Introduction	1
1.1 Definition of the Problem	1
1.2 Criteria for this Computer Assisted Language Analysis System	2
1.2.1 Rapid Processing	2
1.2.2 Generalizability	3
1.2.3 Reliability	3
1.2.4 Applicability to other Languages	3
2. Conceptualizing Natural Language	3
2.1 A Sample Conversation	3
2.2 Information Exchange	6
2.3 A Model of Interaction	9
2.4 Conversation as an Index	12
3. A Case Grammar View of Language	16
4. Processing of Natural Language Data	22
4.1 Automatic Processing of Natural Language Text	22
4.2 The First Phase of CALAS: Grammatical Class Assignment Based on Function Words	24
4.2.1 Operation of MYRA	25
4.2.2 The Dictionary	25
4.2.3 Illustration of the Operation of MYRA	25
4.2.4 Conclusion	33
4.3 The Second Phase of CALAS: Phrase Grouping Process	34
4.3.1 Exo-Relational Phrases	35
4.3.2 Endo-Relational Phrases	35
4.3.3 Arelational Phrases	36
4.4 The Third Phase of CALAS: Clause Separation	40
4.4.1 Independent Clause Separation	41
4.4.2 Dependent Clause Partitioning: I Subordinated Clause Separation	42
4.4.3 Dependent Clause Partitioning: II Relative Clause Separation	44
4.4.4 Dependent Clause Partitioning: III Partial Clause Separation	45
4.4.5 Summary for Clause Separation Process	48
4.5 The Fourth Phase of CALAS: Computer Assisted Case Role Assignment	51
4.5.1 Verb Classification	51
4.5.2 Essential Case Assignment	54
4.5.3 Perpheral Case Assignments	54

5. CALAS as a Basis for Interpreting Natural Language 57
5.1 Introduction 57
5.2 Potential Information Units 57
5.2.1 Verb Types 59
5.2.2 Use of Essential Cases with each Verb Type 59
5.2.3 Use of Peripheral Cases with each Verb Type 61
5.2.4 Comparisons of Speakers' Use of Potential Information Units 61
5.2.5 Lexicon and Use of Potential Information Units 64
5.2.6 Sequence and Use of Potential Information Units 64
5.2.7 Summary 64

6. Future Plans 64
6.1 Further Refinement of CALAS 64
6.2 Empirical Use of CALAS 67

References 70

LIST OF FIGURES

2. Conceptualizing Natural Language

2.1 Spectator's View of Two-Party Interaction Via Natural Language 10
2.2 Sample Dialogue, View One: Party 1, Experience Space of Party 1 and Informative Display 1 13
2.3 Sample Dialogue, View Two: Party 1, Experience Space of Party 1, Prior Overlap, Party 2, Experience Space of Party 2, Informative Display 2, Subsequent Overlap, Common Understanding. 14

3. A Case Grammar View of Language

3.1 Verb Classifications for CALAS 19
3.2 Case Roles Associated with each Verb Type 21
3.2 Case Role Assignments for Sample Dialogue 23

4. Processing of Natural Language Data

4.1 Flowchart of MYRA 26
4.2 The MYRA Dictionary 28
4.3 Flowchart for Phase Grouping Algorithm 37
4.4 Flowchart for Case Assignment Program 52
4.5 Verb Types and Case Roles 53
4.6 Prepositions, Adverbials and Peripheral Case Roles 55

5. CALAS As a Basis for Interpreting Natural Language

5.1 Potential Information Units from Natural Language 53
5.2 Verb Types from Sample Dialogue 60
5.3 Essential Case Use in Sample Dialogue 62
5.4 Peripheral Case Use in Sample Dialogue 63
5.5 Speakers' Profiles from Sample Dialogue 65

6. Future Plans

6.1 Case Assignment of Dependent Clauses and Phrases 68

LIST OF TABLES

4. Processing of Natural Language Data

4.1 Phrase Catagories 35

1. Introduction

1.1 Definition of the Problem

We derive a large portion of what we "know" from the written and spoken language of others. In turn, written and spoken language are the primary means through which we express our thoughts, question others, and share our experiences. Interaction with others can serve many purposes. Some of the more common ones include: describing events, evoking memories, influencing behavior, and searching for agreement.

While language and communication have been studied and debated for many years, much of what happens when humans communicate in verbal, face-to-face conversation is unknown. Information scientists, linguistists, and others have devoted considerable effort to analyzing language patterns, but still very little is known about the subtle transition between language as a collection of data and language as a source of information. Although we may be familiar with the structure of an individual's speach, the lexical items he uses, and how this combination conforms to "an approved grammar" of a particular language, we still cannot explain how he uses the language to achieve different purposes. We can record conversation verbatim, but it is not yet clear how different speech patterns produce agreements, conflicts, influence, or other outcomes.

The research reported in this document is directed toward the following question: Can we determine and consequently predict outcomes of human interaction via conversation, if we are aware of the pattern of speech presented by the participants? To deal empirically with this question requires consideration of a prior question: Can speech patterns which contain information indicative of what is happening in the conversation be detected, isolated, and displayed? If this can be accomplished in a reliable, efficient manner, then the major question of predicting interaction outcomes can be addressed. This report focuses upon the prior question, and describes a reliable, computer-assisted method for isolating and displaying language utterances found in conversation. We believe this method will be helpful to those who wish to ask questions about human interaction via face-to-face conversation.

This report is divided into five sections. This section (Section 1) contains an introduction to the research and the criteria used in developing the system of language analysis described subsequently. Section 2 discusses conceptualizations of natural language interaction through the presentation of

(1) a dialogue and a brief analysis of it
(2) a conceptualization of the dialogue
(3) a model for language interaction, and
(4) a conceptualization of conversation as index.

Section 3 presents a description of the method adopted for the representation of language presented in interaction. This method is a modified form of case grammar which is based on the work of Fillmore (1)

Chafe (2) and Cook (3) but differs somewhat from their contributions. Section 4 presents the mechanics of automated processing of natural language. This section consists of five sub-sections: (1) Grammatical Class Assignment, (2) Phrase Grouping Process, (3) Clause Separation, (4) Case Role Assignment and (5) Language Sorts and Displays. Each of these subsections presents a major phase of the automated language analysis system.

1.2 Criteria for the Language Analysis System

Several criteria have guided the development of this Computer Assisted Language Analysis System (CALAS). These criteria include: (1) rapid processing, (2) generalizability, (3) reliability and (4) applicability to other languages. These criteria are discussed below.

1.2.1 Rapid Processing When one wishes to interpret dialogue or language text, it is important that the entire document be surveyed so that none of the significant features of the text will be omitted from consideration. The hazards of misinterpretation are always great in language communication, but these hazards are markedly increased when one takes a statement or two from context and makes his interpretation of a document from these isolated statements rather than from a survey of the entire discourse. In document search one quickly discovers that language consists of more than lexical items. While lexicon is important, the sequence of and the context in which the words are presented provide the reader or listener with additional clues from which he infers what the speaker intends. The following sentences are illustrative of this point.

(a) The hunter killed the bear.
(b) The bear killed the hunter.

To differentiate event (a) and event (b) the speaker and listener each use data provided by the order in which the words are presented. Structure however cannot provide all the data needed to infer what a speaker intends. For example,

(a) The patient left the operating room in good condition.
(b) The janitor left the operating room in good condition.

The inferences one derives from each of these sentences are likely to differ although the sentence structures are identical. And without more data it is impossible to state whether it is the patient, janitor or operating room that is in good condition.

Surveying an entire document by hand, making an accurate record of the structural patterns and lexical entries is very difficult. Computer assistance is needed for this task. CALAS has therefore been developed to process natural language data rapidly with no manual pre-processing or off-line operations. Natural language is key punched verbatim from transcripts or texts, and from this point on the processing is completely automated, a fact that permits a large amount of data to be analyzed in short order.

1.2.2 Generalizability CALAS is intended for use with many areas of Engilish language discourse. For this reason, procedures were required which depend minimally upon dictionaries. If automatic processing is highly dependent on dictionary entries, then the system is only useful for those dialogues that contain the specified lexical items. The dictionary for CALAS is quite small, containing fewer than 500 words. Many of these words are function words, such as prepositions and articles, which are not topic specific. Thus, processing depends largely on structural analysis cues rather than on dictionary look up. In so far as possible, processing instructions are based upon sequence and context. This allows the processing algorithms to be applied not only to many areas of English language discourse, but also to language usage that may not be in an "approved" or "correct" form. Processing need not be stopped for non-conventional usage. This is an important feature of the system as communication does not always occur through "approved" or "correct" language.

1.2.3 Reliability Any measurement must be reliable if it is to be useful. So in any language analysis system it is important that the same procedures always yield the same results. How users choose to interpret results, of course, may vary. But the more individual judgments that are eliminated the more reliable are the measurements that are produced by the system. A number of individual judgments were required to establish CALAS, but these are based on a description of language which results in errors that are systematic rather than capricious. And no individual judgments are made during processing so the user cannot deliberately or inadvertently influence the results by applying the processing rules non-systematically, or in other ways producing non-uniform interpretations of the data.

1.2.4. Applicability to other Languages CALAS was developed for use with the English language but work has begun on developing a system for Spanish-language analysis and it is believed such systems may be successful with many languages. Each language requires that the system contain some processing rules unique to that language, but the principles of case-role analysis appear to be applicable to a number of languages. Since output from analysis of different languages would be similar, there would be a basis for comparison of results across languages. And such comparisons can be useful in cross-cultural research. The criteria for developing CALAS: (1) rapid processing, (2) reliability, (3) generalizability and (4) applicability to other languages, make it more likely that the system can be used in a variety of communication research projects.

2. Conceptualizing Natural Language

2.1 A Sample Conversation

The question under investigation is: Can speech patterns which contain information indictive of what is happening in a conversation be detected, isolated and displayed? The investigation began by surveying many examples of language dialogue and text. Whether one participates

in or observes a conversation, he is constantly forced to make inferences about what the speaker intends. In surveying texts, efforts were made to isolate structural patterns that might provide information as to how participants and observers make these inferences. For example, consider the following conversation.

HUSBAND: "Dana succeeded in puting a penny in a parking meter today without being picked up.

WIFE: Did you take him to the record store?

HUSBAND: No, to the shoe repair shop.

WIFE: What for?

HUSBAND: I got some new shoe laces for my shoes.

WIFE: Your loafers need new heels badly." (Garfinkel (4))

In observing this conversation one interprets what is happening by combining the language presented with his knowledge of the situation and his knowledge of the participants who produced the exchange. If such an interpretation is to provide an accurate representation of the dialogue, then the observer must (1) correctly infer what the participants take for granted about each other and (2) what they agree on before, during and after the conversation. This is difficult to do. The following is one observer's account of what each participant believed to be happening as the conversation progressed.

HUSBAND:	Dana succeeded in putting a penny in a parking meter today without being picked up.	This afternoon as I was bringing Dana, our four-year-old son, home from the nursery school, he succeed in reaching high enough to put a penny in a parking meter when we parked in a metered zone, whereas before he had to be lifted to reach that high.
WIFE:	Did you take him to the record store?	Since he put a penny in a meter that means that you stopped while he was with you. I know that you stopped either on the way to get him or on the way back. Was it on the way back, so that he was with you or did you stop there on the way to get him and somewhere else on the way back?
HUSBAND:	No, to the shoe repair shop.	No, I stopped at the record store on the way to get him and stopped at the shoe repair shop on the way home, when he was with me.
WIFE:	What for?	I know of one reason why you might have stopped at the shoe repair shop. Why did you in fact?

HUSBAND:	I got some new shoe laces for my shoes.	As you will remember I broke a shoe lace on one of my brown oxfords the other day so I stopped to get some new laces.
WIFE:	Your loafers need new heels badly.	Something else you could have gotten that I was thinking of. You could have taken in your black loafers which need heels badly. You'd better get them taken care of pretty soon.

This account from an observer is helpful for others who wish to interpret the exchange, but it is cumbersome and much longer than the original exchange. And although the explanation is lengthy, it is still incomplete. Further explanation is probably needed (an explanation of the explanation) for a number of observers to agree on what each speaker intends and what each infers about the other's intentions. And, the problem of inference remains whether the individual interpreting the conversation is a participant or an observer; it even remains when both participants are interpreting the dialogue to others or to each other. The task of interpreting a conversation by further explanation becomes increasingly difficult and cumbersome, if not impossible. Explaining the explanations becomes a never ending sequence and inferences as to what speakers intend often become more difficult to make as the explanations pile up.

This brief example of dialogue and interpretation illustrates the problem involved in analyzing natural language and the difficulties involved in isolating the structural cues observers use to make inferences about conversation. These problems of interpretation exist even when participants and observers are native speakers of a language and declare that they are strongly motivated to understand one another.

In initial surveys of dialogue, such as that above, efforts were made to determine what factors might be helpful in answering questions such as the following.

> Why was the wife immediately able to question the <u>place</u> of the event (the event being Dana's success at reaching the parking meter) without requiring more explanation of the process of the event itself?
>
> In turn, what signals did the wife produce that allowed the husband to shift the discussion easily from process to place? How can an observer infer what signals are important in the interchange particularly if he has no prior knowledge of the participants or the situation?
>
> In short, what allows the exchange to proceed as if the participants understood one another?

Some intuitive analysis provides an observer with more data as to what is happening. For example, the initial comment by the husband describing the event contains four phrases which seem to make the explanation of the event more explicit.

"Dana succeeded

(1) in putting a penny
(2) in a parking meter
(3) today
(4) without being picked up."

From the above output presented by the husband, the observer, without prior knowledge of Dana or the participants, becomes informed about:

(1) the nature of the event (placing a penny)
(2) the location of the event (in a parking meter)
(3) the time of the event (today)
(4) the manner of the event (without being picked up).

The wife's response, which is a query asking for more explicitness about the event, elicits from the husband a language string that has similar elements to the language string the wife presented in her question.

Compare the last part of each of the word strings:

Wife: to the record store?

Husband: to the shoe repair shop.

Each of these phrases signals more specificity about the location through the use of the function words, to the. Once the participants had exchanged these language strings, they proceeded to exchange signals about a topic (the reason for going to the shoe shop) that was different from the topic (Dana's actions) which was initially presented.

While this partial intuitive analysis may provide the observer with some notions about patterns present in this dialogue, it raises more questions than it answers about what makes such exchanges possible. Initial survey and intuitive analysis of other documents raised many such questions and further indicated the need for a reliable systematic method of language analysis. This need is shared by research specialists in social science, information science, education and by others who are concerned about what transpires in human interactions.

2.2 Information Exchange

From the sample conversation presented above one can infer that "something" is being exchanged by the participants and that this "something" is being signaled or signified through the natural language each participant uses. The major portion of the research to date has focused upon analyzing these signals to determine their structure and those attributes of that structure which speakers and listeners use to infer what is going on. We have termed such signals informative displays. For example, what in the husband's initial language display signaled to the wife or prompted the questioning response: "Did you take him to the record store?" If an observer knew more about the nature of such signaling, he might be able to predict outcomes such as whether participants believed they understood one another well enough to engage in conjoint activities.

We have labelled such understanding common understanding and such conjoint activities concerted actions. As the discussion proceeds the usage of the terms informative display, common understanding, and concerted actions will become more evident when they are presented as variables in a model of interaction. The model which is discussed in Section 2.3, presents a characterization of what happens in conversation and based on a particular conceptualization of information exchange. The term "information exchange" refers to the "something" that is transmitted and to the process that allows that "something" to pass back and forth between participants. Therefore before studying the model, a few remarks about information exchange are in order.

Since the focus of this research is upon interaction, humans or other systems that are to exist in total isolation from their environments are outside our realm of consideration. In system/environment interaction, an observer can record data the system receives from the environment and data that the system emits to its environment. In the example above, let us term the wife the system, and the husband an element of her environment. The system (wife) receives the following data from the environment (i.e., the husband; we shall ignore for the moment all other aspects of the wife's environment).

> "Dana succeeded in putting a penny in a parking meter today without being picked up."

Upon reception of this data, the system (the wife) produces the following output or observable:

> "Did you take him to the record store?"

The observer can also record the conditions of the environment at the time of interaction; however, he can only infer what data become information for the system, and what possible courses of action, based on this information, the system has identified.

In our example, then, the husbsnd begins the interaction in a state associated with the production of language strings. When he changes to a different state, the system (wife) emits a language string. This process of language on, language off on the part of both the husband and the wife continues throughout the interaction. Until the wife (system) emits an observable ("Did you take him to the record store?"), the observer has no idea what data in the language string emitted by the husband, have provided the wife with information. Her response indicates that she is acting as if she understood most of the "something" contained in the signals of her spouse. Different inferences would be in order if the wife had responded: "How's come", "I don't understand", "Mildred was over for lunch today", or any other of the myriad responses she could have produced.

Prior to interaction, a system (in this case the wife) has no direct data on the state of the environment (in this case the husband). The environment could be in any of a number of different states. For example, the husband might have opened the conversation with: "You are

a rotten cook!" Data received by a system from the environment partitions the possible different states into: (1) those states that have been observed, and (2) those states that have not been observed. Therefore, the reception of data by the system yields a reduction in the size of the set of a priori inferences concerning alternative states of the environment. This reduction of the number of alternatives, or increases in certainty about the environment, is information in the classical Maxwell/Boltzmann sense.

For example, when the husband produced his initial language string: "Dana succeeded in putting a penny in a parking meter today without being picked up." the set of inferences that the wife might make about the state of the environment was greatly reduced (though perhaps still large). After this initial statement the wife could observe that the topic of conversation was Dana, and not her cooking, or her husband's day at work. She could also observe that his statement was not an inquiry to which she was expected to respond by supplying information. In short the initial statement presented to the system substantially reduced, for the system, the number of alternative states of the environment.

Information received by a system in interaction with its environment may be used by it to select a course of action. The range of possible courses of action is limited by the attributes of the system under consideration. Under the appropriate conditions, the execution of a course of action can yield observables that can be recorded (measured) by a spectator and/or by the system itself. After a particular action has been exhibited by the system, its consequence can be inferred by observing subsequent response(s) (changes in signals emitted) from the environment. For instance at the end of the conversation, the wife might have suggested returning to the shoe store to have heels put on her husband's shoes. If the husband had received such a suggestion from the wife, he might have agreed and gone, asked her to go, said "No" and not gone, or said nothing. Change in informative display foom the environment (in this case the husband).is measured by comparison of these subsequent responses (new data) with the initial ones received by the system. This comparison is feedback, as defined in the Yovits/Ernst description of information exchange (5).

Choice of course of action and its execution can be viewed as the testing of a hypothesis concerning the state of the environment. Data received by the system can also serve to test a particular hypothesis about the environment.

Thus, the wife, when she asked "What for," (as a response to her husband's signal "No, to the shoe repair shop."), might have been testing whether her husband had his loafers fixed. Of course, based on this portion of the dialogue alone, the observer might infer that the wife had no idea why her husband went to the shoe repair shop. However, her later comment "Your loafers need new heels badly" gives support to the first inference that the wife was checking to see if her husband had the heels of his loafers replaced. Frequently, a new hypothesis about the disposition of the environment will be formulated based on the result of the feedback from observables. The data received by the

system from the environment are informative if they serve to : 1) test a hypothesis and 2) permit decision making about which course of action to follow.

2.3 A Model of Interaction

The interactions under study in this research are those which are managed via natural language. People are observed to do things together, to negotiate agreements, to sign contracts, and the like. As indicated in the discussion concerning the conversation presented above, we assume that one means of accomplishing these conjoint activities is through conversation, and that, in conversation, participants signal to one another their interpretations of and intentions toward what is happening or what is expected to happen.

As mentioned above, we have labelled these signals informative displays. While we have a label for such signals we are not yet able to describe adequately all of their attributes. But through CALAS we have isolated language components which we believe to be useful for identifying and characterizing these informative displays in natural language. For now, inquiry is limited to the language itself, exclusive of extra-linguistic signals, and as mentioned above centers upon the structural or syntactical component of language. Extra-linguistic signals, e.g., loudness, tone, gestures, and the like, although interesting to consider, are presently beyond the scope of our work.

In order to explore the nature of these natural language signals we have studied the context(s) in which they appear. From such study a model of interaction has been developed which is amenable to analysis component by component. While the model (see Figure 1) is tentative, it has enabled us to characterize the kinds of interaction we are studying. The model represents a two-party interaction (such as found in our husband-and-wife example). These parties may be humans, machines, or one of each. Although we are primarily concerned with natural language and therefore with human behavior, the model itself applies equally well to a human-human, human-machine or machine-machine interface.

From the model, it seems evident that the participants' definitions of the situation are a central feature in characterizing any interaction. Thus, we are interested not only in the signals that each participant provides, but also in the interpretations each imposes on those signals. The model itself, however, is to be interpreted from the point of view of a non-participating spectator.

In constructing the model, the following assumptions have been made.

1) That each participant enters the interaction with a unique experience space.
   a) An experience space is a construct that we find helpful to account for the likelihood that participant behavior is not random.

2) That this experience space may include preconceptions of what will occur in the interaction.

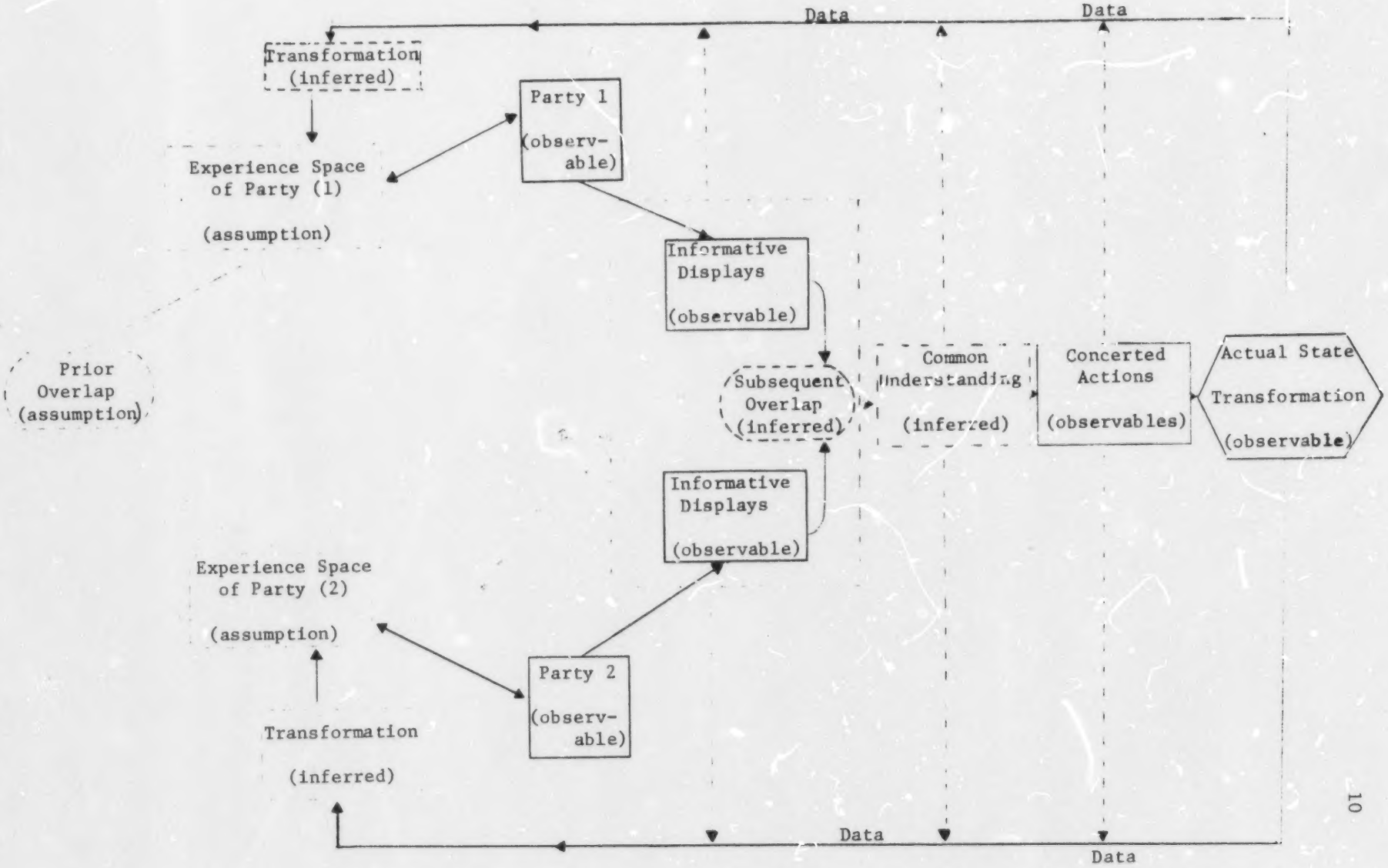


Figure 2.1 Spectator's View of Two-Party Interaction via Natural Language

3) That neither the spectator nor the participants are able to observe directly but must infer the existence, structure and content of these spaces (ES_1) and (ES_2), and

4) That prior to the interaction, overlap (interaction between ES_1 and ES_2 exists, but cannot be measured directly; prior overlap can only be inferred from data obtained before, during or after the interaction.
   a) Such overlap may, in principle, be null.

The model demands that the observer treat the verbal output of each participant as something that can be partitioned into and interpreted as informative display. A non-participating spectator may begin his observations at any point in the sequence of events depicted in the model. The interactions with which we are now concerned, however, involve reciprocated signaling from one participant to the other. The exchange of signals as informative displays allows the participants and/or the spectator to draw inferences about the experience spaces of the participants. This exchange may result in an increased overlap of experience spaces. When the overlap increases, shared agreement or "common understanding" may be achieved, but, in principle, common understanding may not occur.

As informative displays are presented accompanying changes in the experience spaces of the participants also occur. This point is illustrated in the model by means of the feedback paths to the experience spaces of Party 1 and Party 2. Data from informative displaying have the potential of being transformed into information by the participants as they proceed in the interaction. The model assumes that this process of informative displaying and feedback is recursive. Again, feedback and information are used here as defined by Yovits and Ernst (5).

At some point in time, a level of common understanding may be reached that permits the participants to initiate some process or task that requires their concerted action(s). Such concerted action results in some observable event(s) or outcome(s). We emphasize that this outcome may be observed and evaluated in terms of <u>whatever</u> criteria one or both participants or the spectator may choose to impose.

Although the model generates a number of questions we are as yet unable to answer, it resolves many of the difficulties inherent in attempts to characterize interaction and conjoint activity. For example, the model allows for all possible alternative goal specifications:

(a) both participants have the same goal
(b) each participant has a different goal
(c) neither has any definite goal
(d) one participant has a goal; one does not
(e) the goal is externally imposed (i.e., by something or someone other than the participants)
(f) the goal(s) change during the operation of the model
(g) the goal(s) develop during the operation of the model

Every component of this model effects a change in the experience space of each participant and the constantly changing experience spaces provide the flexibility needed to encompass the aforementioned goal alternatives.

Since the model operates equally well whatever the goal of the interaction, the observer is freed from having to account for the initial assumptions of the participants. Thus the observer can concentrate on what is presented in the interaction and draw his inferences from this rather than from his preconceived notions of what was "supposed" to occur. While in some instances prior knowledge of the participants and of the situation can assist the observer, we are interested in the inferences an observer can make from the dialogue itself. And thus the model focuses on the interaction itself.

This model has enabled us to isolate and investigate in detail what we believe to be the essential components of interaction. These components are: the assumptions of experience space, prior overlap and subsequent overlap; the inferences of common understanding and transformation; and the observables of parties, informative displays, concerted actions, and actual state transformations. The parties can be symbolically represented by Speaker 1, Speaker 2, husband, wife, etc. Since extra-linguistic signals are beyond the present scope of the work an observer can work from a verbatim record of the interaction in which the speakers are clearly differentiated. The model does not require that the observer be present at the interaction but only that he have a verbatim record of it.

While all the components of the model are interesting subjects for study, the current focus of the investigation, and the principal topics of this technical report are a description of the component informative display, and the development of a computer assisted method of identifying and classifying its attributes. Informative display was the first component of the model chosen for study because, although we cannot delineate the attributes of informative displays, we can observe and partition the language strings that contain signals (e.g., words and combinations of words in everyday English usage). People do speak words in sequence and we can record these words and sequences verbatim. Informative display was also chosen because it is believed to be an independent variable (the manipulation of signals in interaction will effect both common understanding and concerted action). This variable needs to be isolated before other concepts specified by the model can be characterized. To say that informative displays are signals exchanged by participants in an interaction is not a sufficient definition. One needs to identify the specific attributes of these signals; how these attributes vary in context, and whether different kinds of displays produce different outcomes. Figures 2.2 and 2.3 illustrate how the sample dialogue is viewed in terms of the model.

2.4 Conversation as an Index

After conceptualizing what we assume to occur in human interaction we once again surveyed natural language text for systematic, formal ways to partition the dialogue. Using extensive empirical observations, and the formal characterization of interaction provided by the model as

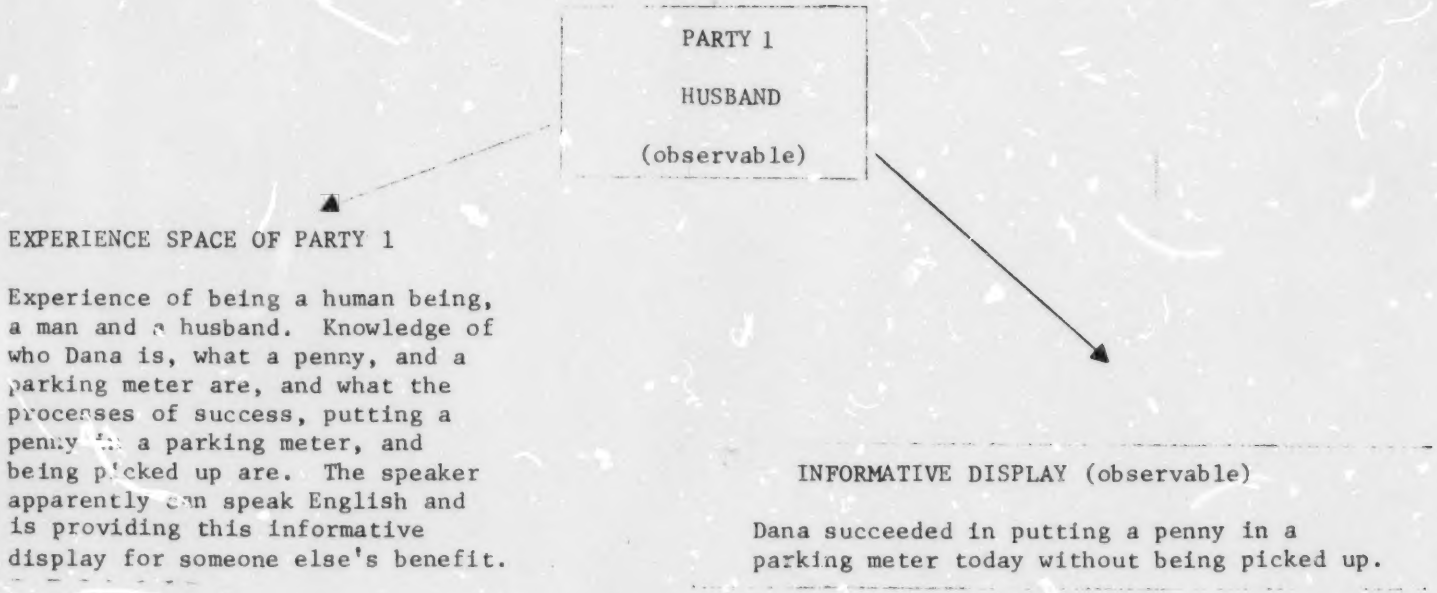

PARTY 1
HUSBAND
(observable)
EXPERIENCE SPACE OF PARTY 1
Experience of being a human being, a man and a husband. Knowledge of who Dana is, what a penny, and a parking meter are, and what the processes of success, putting a penny in a parking meter, and being picked up are. The speaker apparently can speak English and is providing this informative display for someone else's benefit.
INFORMATIVE DISPLAY (observable)
Dana succeeded in putting a penny in a parking meter today without being picked up.

Figure 2.2 Sample Dialogue, View One: Party 1, Experience Space of Party 1 and Informative Display 1.

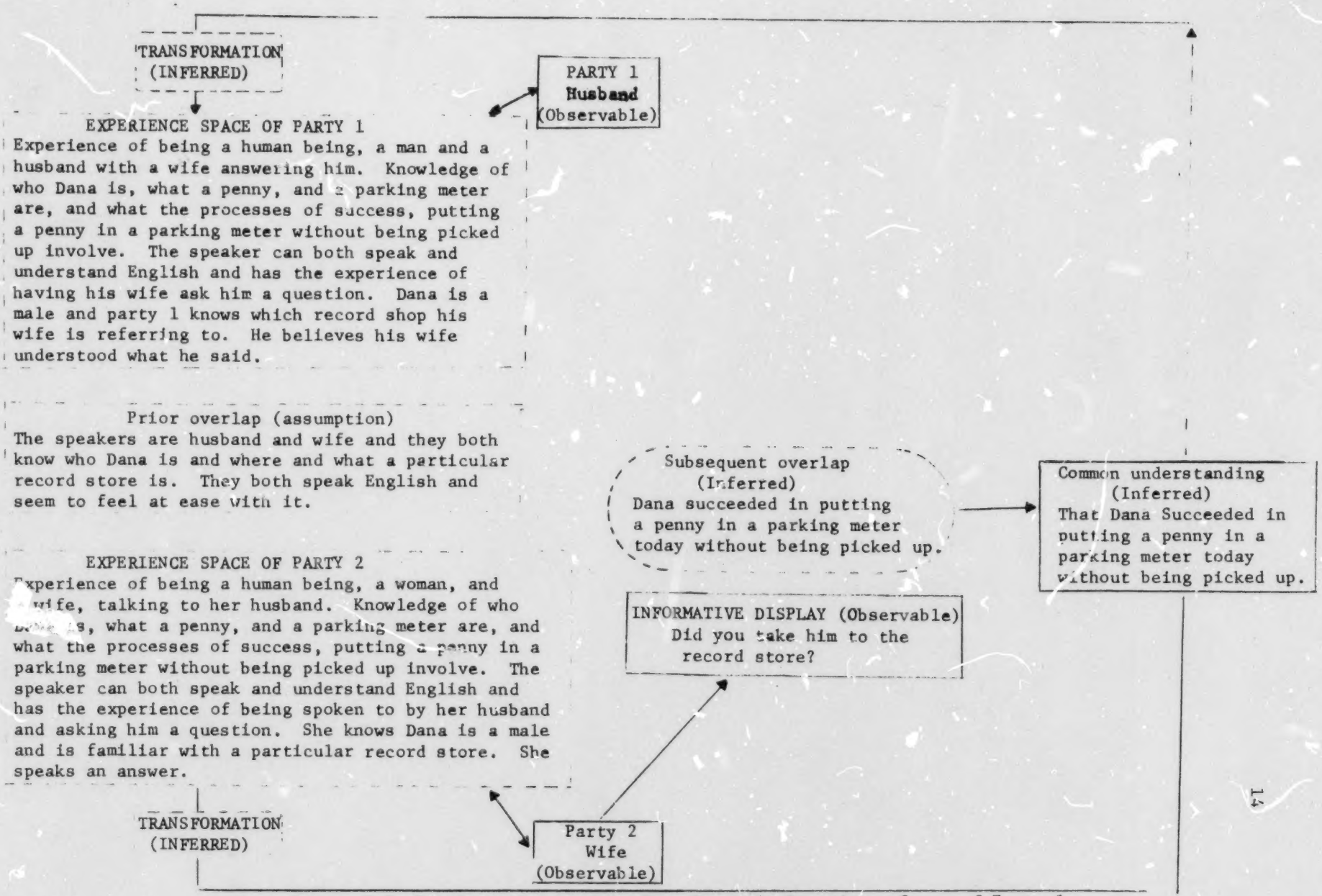

TRANSFORMATION
(INFERRED)
PARTY 1
Husband
(Observable)
EXPERIENCE SPACE OF PARTY 1
Experience of being a human being, a man and a husband with a wife answering him. Knowledge of who Dana is, what a penny, and a parking meter are, and what the processes of success, putting a penny in a parking meter without being picked up involve. The speaker can both speak and understand English and has the experience of having his wife ask him a question. Dana is a male and party 1 knows which record shop his wife is referring to. He believes his wife understood what he said.
Prior overlap (assumption)
The speakers are husband and wife and they both know who Dana is and where and what a particular record store is. They both speak English and seem to feel at ease with it.
EXPERIENCE SPACE OF PARTY 2
Experience of being a human being, a woman, and a wife, talking to her husband. Knowledge of who Dana is, what a penny, and a parking meter are, and what the processes of success, putting a penny in a parking meter without being picked up involve. The speaker can both speak and understand English and has the experience of being spoken to by her husband and asking him a question. She knows Dana is a male and is familiar with a particular record store. She speaks an answer.
TRANSFORMATION
(INFERRED)
Party 2
Wife
(Observable)
INFORMATIVE DISPLAY (Observable)
Did you take him to the record store?
Subsequent overlap
(Inferred)
Dana succeeded in putting a penny in a parking meter today without being picked up.
Common understanding
(Inferred)
That Dana Succeeded in putting a penny in a parking meter today without being picked up.

Figure 2.3 Sample Dialogue, View Two: Party 1, Experience Space of Party 1 Prior Overlap, Party 2, Experience Space of Party 2, Informative Display 2, Subsequent Overlap, Common Understanding.

a basis, we began by partitioning sentences into fragments. Certain fragmentation points, such as prepositions and conjunctions, were established, and sentences were partitioned at these points. The results seemed unconnected, and not representative of what the participants presented in the interaction.

A different approach was required to preserve the sense of order, and the expression of units that are apparent in natural language. Realizing that a unit was needed which would make the language signals within it easily amenable to analysis, we became interested in the possibilities of the grammatical clause as such a unit. A clause is a string of words that contains one and only one predicate (6). We partitioned natural language dialogues into clauses and then attempted to describe each clause in terms of a topic and a comment. The topic was considered the subject of the predication (in the topical, not the traditional, grammatical sense), and the comment, what was said about that topic. This partitioning better described the interaction than the phrase-fragmentation approach but did not provide a means of description for every element in the clause. Many language strings seemed to fit neither the topic nor the comment category; therefore, signals presented by the participants in the interaction were being lost. These two approaches were thus abandoned.

We then approached the problem of analyzing language strings as one of index creation and document retrieval. It became apparent that the analysis of person-to-person interaction via natural language is similar to, if not identical with, the analysis and subsequent indexing of written documents. The goal of each is total representation and accurate retrieval of the document content. To ensure accurate representation and retrieval, a data element and its associated relations need to be identified. The goals of communication analysis and indexing are seldom realized in actual practice. There are many examples of speakers "clarifying their remarks", or claiming they have been misrepresented; similarly most indexes do not provide adequate information as to a document's contents (7). These failures accentuate the need for consistent procedures for developing interfaces between documents and their users. Such procedures can also be used for the development of a system of language analysis to make what the speaker "puts out there" more useful to participants and/or observers.

Landry has proposed that for a given data element the indexing process represents the following items (i.e., provides the following structure):

(a) the data element itself (symbol)
(b) the surrounding data elements (context)
(c) the order of surrounding data elements (syntax)
(d) relations to other data elements (semantics) (8).

The indexing process yields an index that may be viewed as an interface between a document and a potential user. The index provides the potential user of the document with a measure of its contents:

> "In a formal sense both indexing and measurement involve the production of a result from a classificatory act on an object of interest" (8).

In this context then, to analyze and represent human interaction accurately requires an index interface which accurately represents the four attributes of structure mentioned above: symbol, context syntax and semantics.

In viewing language analysis within an indexing framework, we have maintained the grammatical clause as the unit for further analysis, and have turned to case grammar as the representational interface between the language signals of the participants, and the inferences derived from this language by observers.* Case grammar provides a means of viewing each data element in ordered context, and in relation to other data elements presented in the interaction. Case grammar also provides a means for representing natural language and is the perspective from which the computer automated language analysis system (CALAS) was developed.

3. A Case Grammar View of Language

Case grammar characterizes the structure of language through a description of relationships among components of that structure. Case grammar relationships: (a) are derived directly from the language strings presented in the interaction, (b) are not constrained to specific situations or topics, and (c) can be applied to a number of languages. Therefore a case grammar representation meets three of the criteria for a language analysis system presented in Section 1.2. The criterion of rapid processing will be discussed later.

Our treatment of case grammar differs from the one presented by Fillmore (9, 10, 11). It builds upon Fillmore's work, but includes the conceptualizations of Chafe (2), Cook (3), and work completed by others (3, 12-19). Case grammar proposes that grammatical structure consists of a series of non-linearly ordered, case-marked noun phrases associated with a verb phrase (1). Cases may be viewed as roles which retain their character while participating in different natural language utterances. The verb phrase is the pivotal word class in language analysis via case grammar. The verb phrase is surrounded by noun phrases. The noun phrases exhibit certain relationships to the verb phrase and thus to each other and to the remaining phrases in the sentence. Every phrase but the verb phrase is a case candidate, and unless the phrase is embedded within another phrase, it performs a role or function

---

* Theoretically, an observer can be participant or spectator, but for now the interaction is being studied from a spectator's point of view.

within the clause, and consequently is given a case designation. Case designations are of two kinds. First, there are essential cases which are governed by the kind of verb found in the clause; these are sometimes called nuclear or propositional cases. Second, there are peripheral cases, also called modal cases, which occur frequently with a wide variety of utterances but are not governed by the type of verb contained in an utterance.

For example:

(a) The hunter killed the bear.
(b) The bear killed the hunter.

In each of these examples there are two essential and no peripheral cases. The pivotal word class is killed. In example (a) the noun phrase the hunter takes the role of the actor (agent) and the bear takes the role of the acted upon (object). In example (b) the case roles are reversed. The order of the word string and the nature of the verb provide the necessary information for describing case roles.

(a) He won with ease.
(b) In the villages the men worked for money.

Example (a) contains two case roles: he is the actor (essential case), and with ease (peripheral case) tells us something about the manner of his action. Example (b) contains one major case: the men, as actors (or agents) and two peripheral cases: in the villages and for money. The first gives the reader some idea about location (locative) and the second, some idea about the reason for the action (causative). Fillmore (1) and Chafe (2) both postulate that language structure can be characterized as consisting of a verb phrase and a series of noun phrases. But they disagree as to the centrality of the verb.

> "According to Fillmore, the sentence consists of a verb and one or more noun phrases, each associated with the verb in a particular case relationship. But although the noun cases are related to the verb, it is the noun that selects the verbs and not vice versa.'The verbs are selected according to the case environments which the sentence provides--what I shall refer to as the case frame,' (Fillmore, 1). Fillmore's point of view seems to be that there are pre-existing case frames, into which verbs are inserted and further 'many verbs are capable of occurring in more than one case environment.'"

> "According to Chafe (2), however, the typical configuration is that of a central verb, accompanied by one or more nouns each of which stands in some particular...relation to the verb. In these configurations he says 'the verb will be assumed to be central and the nouns peripheral.'"(3)

Chafe's position seems to be the stronger one, as a noun phrase does not assume the properties of a case role or perform a function until it is placed in context with a verb phrase. According to

Chafe's view, then, it is the verb that determines the case roles that may surround it, and not the relationship among roles that determines the verb. This position makes clear that verb classification is an important prerequisite to accurate case role assignment.

The verb classifications used in this research are based on the classifications proposed by Chafe (2) and modified by others (12,16, 19). These classifications have been further modified to make them more amenable to computer analysis. These classifications are: (1) state, (2) benefactive, (3) experiencer and (4) agentive (see Figure 3.1). Section 6.2 of this report discusses the further work that is planned in classifying verbs.

A state verb is any form of the verb "to be" when that verb is the main verb in the clause. A benefactive verb is any form of the verb "to have" when that verb is the main verb in the clause. Experiencer verbs are expressions of feeling, sensing, or knowing. All other verbs are agentive.

For the purposes of the current work in developing CALAS, four essential and six peripheral cases have been designated. Essential cases are limited to those few essential cases found to be necessary to describe the case environment governed by the verb; all other cases are peripheral.

The essential cases are:

A - Agent, the typically animate instigator of the action described by the verb. Agents may be inanimate, where the inanimate noun is presented as if it possessed the potency for instigating action.

E - Experiencer, the typically animate one who experiences the feeling, sensation, etc., described by the main verb. Experiencer is rarely inanimate, but may be, in those cases in which the inanimate object is described as if it were capable of experiencing.

B - Benefactive, the typically animate possessor (in its broadest sense) of some object, whether the possession be temporary or permanent, positive or negative (as, I have a cold).

O - Objective, the typically inanimate receiver of the action described by the main verb; the person or thing being described (with state verbs).

All other cases are peripheral. The essential cases explicated by Fillmore (1) but not used in this analysis are Instrument, Source, Goal. Instrument is considered a subset of the Manner case, Source and Goal as subsets of the Locative case.

Verb Classifications

STATE - Any form of verb to be when that verb is the main verb

e.g.: He was an old man.

She will be the leader.

BENEFACTIVE - Any form of the verb to have when that verb is the main verb

e.g.: I have money.

He will have many choices.

EXPERIENCER - Expressions or feeling, sensing, knowing

e.g.: John loves Mary.

She does not believe him.

AGENTIVE - All other verbs

e.g.: He ran the meeting.

The teacher penalized the absent students.

Figure 3.1 Verb Classifications for CALAS

Peripheral cases are:

L - Locative, the place where the action described by the verb occurs. This includes Source-Locatives, Goal-Locatives, Path-Locatives, or Locatives as place-in-which the action occurs.

T - Time, the time when the action described by the verb occurs. This includes Source-Time, Goal-Time, and Time-in-which, including both Time span and points of Time.

M - Manner, the way in which the action described by the verb is performed. This includes the instrument case as a subset, when the agent is presented; when the agent is absent, the instrument is called an agent.

C - Comitative, the accompaniment case, the typically animate subject accompanying the main actor of the action described by the verb.

Cs- Cause, the cause giving the reason for the action described by the verb. Typically expressed in clauses with because, or phrase with words cause, reason, order, command, etc. Also found in gerund phrases introduced by from.

P - Purpose, the case giving the purpose of the action described by the verb. Typically expressed in clauses with so that. or in order to. Also in phrases with the prepositions for (plus inanimate), and after.

Each essential case is related to a particular verb type in a specific pattern, but the case does not always appear each time the verb is used. However, no other essential cases may appear except those that are related. This relationship is determined by the verb type (see Figure 3.2).

The cases associated with each verb type are as follows:

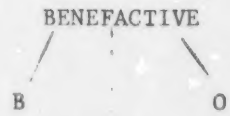
(1) Stative Verb (O case before verb and O case after)
(2) Benefactive Verb (B case before verb and O case after)
(3) Experiencer Berb (E case before verb and O case after)
(4) Agentive Verb (A case before verb, E case (if animate indirect object is included in clause) and O case after the verb).

Peripheral cases are usually assigned to prepositional phrases or single word adverbials. Single word adverbials fill either comitative, locative time or manner case roles. Most often the specific preposition governs the case role which is assigned to prepositional phrases. A detailed discussion of algorithms developed for computer assisted assignment of case roles is presented later in the report. But theoretical positions presented here provided the departure point for the development of CALAS. The algorithms for CALAS are based on the following postulates:

Essential Cases and Verb Types

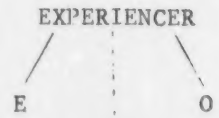
He/ was/ an old man.

She/ will be/ the leader.



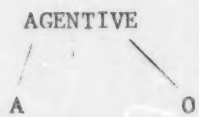
I/ have/ money.

He/ will have/ many choices.



John/ loves/ Mary.

She/ does not believe/ him.



He/ ran/ the meeting.

The teacher/ penalized/ the absent students.

Figure 3.2 Essential Cases Associated with Each Verb Type

(1) the verb phrase is the central word class
(2) the verb has one and only one set of essential cases associated with it
(3) verbs with the same set of essential cases belong to a single verb type
(4) essential case roles are assigned by the verb to the noun phrase
(5) peripheral cases are independent of the particular verb and of the set of essential cases a particular verb requires.

If this system of analysis represents what native speakers of the language "put out there", then classification of verbs recommended by Fillmore (1) and completed for Spanish verbs by McCoy (16) and Aid (19) could be generalized to other languages, and could serve as a basis for essential case role designations.

The reader will recal the sample dialogue introduced at the beginning of this report.

HUSBAND: Dana succeeded in putting a penny in a parking meter today without being picked up.

WIFE: Did you take him to the record store?

HUSBAND: No, to the shoe repair shop.

WIFE: What for?

HUSBAND: I got some new shoe laces for my shoes.

WIFE: Your loafers need new heels badly.

Figure 3.3 shows the case role assignment for this dialogue based on the five postulates above. The following sections present the algorithms used to automate case grammar analysis.

4. Processing of Natural Language Data

4.1 Automatic Processing of Natural Language Text

Once case roles were selected as a means of characterizing natural language, work was begun to develop an automated system to perform rapid and accurate case role assignments. To do this requires automated means for parsing language strings in such a way that algorithms can be developed for computer assisted case role assignments. The parsing procedure has been separated into three distinct operations: (1) grammatical class assignment, (2) phrase grouping and (3) clause separation. After the natural text has been automatically parsed by each of these operations in sequence, it is ready for the fourth phase of processing, case role assignment. The initial phase of this parsing procedure is the determination of the grammatical class of each word in the text. When the research began, such a grammatical class assignment algorithm had already been

A VP	
Dana/ succeeded in	A
VP O L T	
putting/ a penny/ in a parking meter/ today.	O L T
M VP	
without/ being picked up.	M
VP A O L	
Did [you] take/ him/ to the record store?	A O L
L	
No/ to the shoe repair shop.	L
P	
What for?	P
A VP O B	
I/ got/ some new shoe laces/ for my shoes.	A O B
B VP O M	
Your loafers/ need/ new heels/badly.	B O M

Figure 3.3 Case Assignments for Sample Dialogue

developed, and has subsequently been tested and improved so that grammatical class assignment by automatic means is now achieved with a high degree of accuracy and efficiency. The second phase of the parsing procedure uses the data generated by the grammatical class assignment algorithm to divide the data into phrases. The output of the phrase grouping procedure is the input for the clause separation process. And in turn, the output from the clause separation process is input for the final stage, case role assignment. All processing is automated and requires no off-line manual procedures. The next four sections of this report discuss each of these automated language analysis procedures in detail. Together they comprise the Computer Assisted Language Analysis System, CALAS.

4.2 The First Phase of CALAS: Grammatical Class Assignment Based on Function Words

In most automatic language processing systems, there is some phase which requires syntactic analysis. For that reason, a program has been developed which is sufficiently accurate and efficient to be used in current information processing systems. This program, called MYRA (20) has many potential uses, but is presented here as a language parsing component of a system designed to identify and characterize natural language.

There are several reasons why MYRA is superior to grammatical class assignment programs that have developed previously. Until recently, most implementations for syntactic analysis ahve used an approach in which each word in the text is compared against the words in a dictionary. This approach has several limitations. First most dictionaries range in size from 6,000 to 75,000 words (21). A dictionary is expensive both in time and in the amount of storage required. Second, the dictionary may be subject specific, thus limiting accurate grammatical classification to one particular field. In addition, much of the ambiguity as to how to assign specific words is not resolvable by simple dictionary look-up procedures, but must be resolved through context. For example, the word work can function as a noun, adjective or verb depending on the context.

(a) The work is very hard. (noun)
(b) We work everyday. (verb)
(c) Here is his work place. (adjective)

From this simple example it is clear that if syntactic analysis is to be rapid, accurate, and gneralizable, it must be based upon structural characteristics as well as upon dictionary look-up techniques, and to achieve accurate grammatical class assignments, the MYRA program relies heavily on structurual elements in text. We have implemented an algorithm which assigns words in a sentence to their respective grammatical classes based on:

(1) the function words (e.g., prepositions, articles, auxiliary verbs, etc.) and punctuation in a sentence, and
(2) the position of each word in a sentence and the position of each word relative to the surrounding function words.

Only a very small dictionary is used, and the program is fast and relatively small.

There are several advantages to the approach which we have used. First, the size of the dictionary is small, thus keeping processing time low. Second, assignment errors can be isolated; a mistake in the classification of one word does not mean failure in the rest of the sentence. Third, the program will never reject a sentence. At worst, the sentence will be analyzed by a series of defaults. Fourth, by using the structural properties of a sentence most lexical ambiguities such as the one mentioned above are eliminated. Fifth, since this approach does not depend upon large dictionaries, it is applicable to any English text. Most important, however, is that the speed and small size of MYRA make this approach practical for many applications in automatic language processing.

4.2.1 Operation of MYRA MYRA makes three passes through a sentence (see Figure 4.1). In the first pass, each word in the sentence is checked against a dictionary of function words. When a function word is identified, it is replaced by a symbol representing the particular class to which it belongs; otherwise, the word is represented as a blank element. At the end of the first pass, the sentence is represented by an array of elements; each element corresponds to a word in the sentence and specifies the class of the word, if it has been identified. In the second pass, the sentence is again processed sequentially from left to right. As each function word is encountered, rules pertinent to that particular class of function word are applied to the unidentified elements immediately surrounding the function word. The third pass makes assignments for any unidentified elements which were not recognized in the second stage. After all elements in the array have been identified, some elements may then be reassigned, if the overall structure of the sentence is incomplete.

4.2.2 The Dictionary The present dictionary contains 666 words, which are organized into 15 different groups. These groups include auxiliary verbs, conjunctions, pronouns, prepositions, punctuation and determiners (i.e., words such as 'the' and 'a'). A group of more than 300 verbs has also been included in the dictionary. While verbs are not classified as function words, this class was added since verbs were one of the most difficult classes to identify and were a common source of error in earlier versions of MYRA. And the verb is central to accurate case role assignment. Generally, the groups of function words are exhaustive but neither function words nor verbs were included which had a frequency of occurrence of less than .005% in the Kucera and Francis study (22). A partial list of the dictionary is given in Figure 4.2.

4.2.3 Illustration of the Operation of MYRA In illustrating the operation of MYRA, the following notation will be used for the various syntactic classifications:

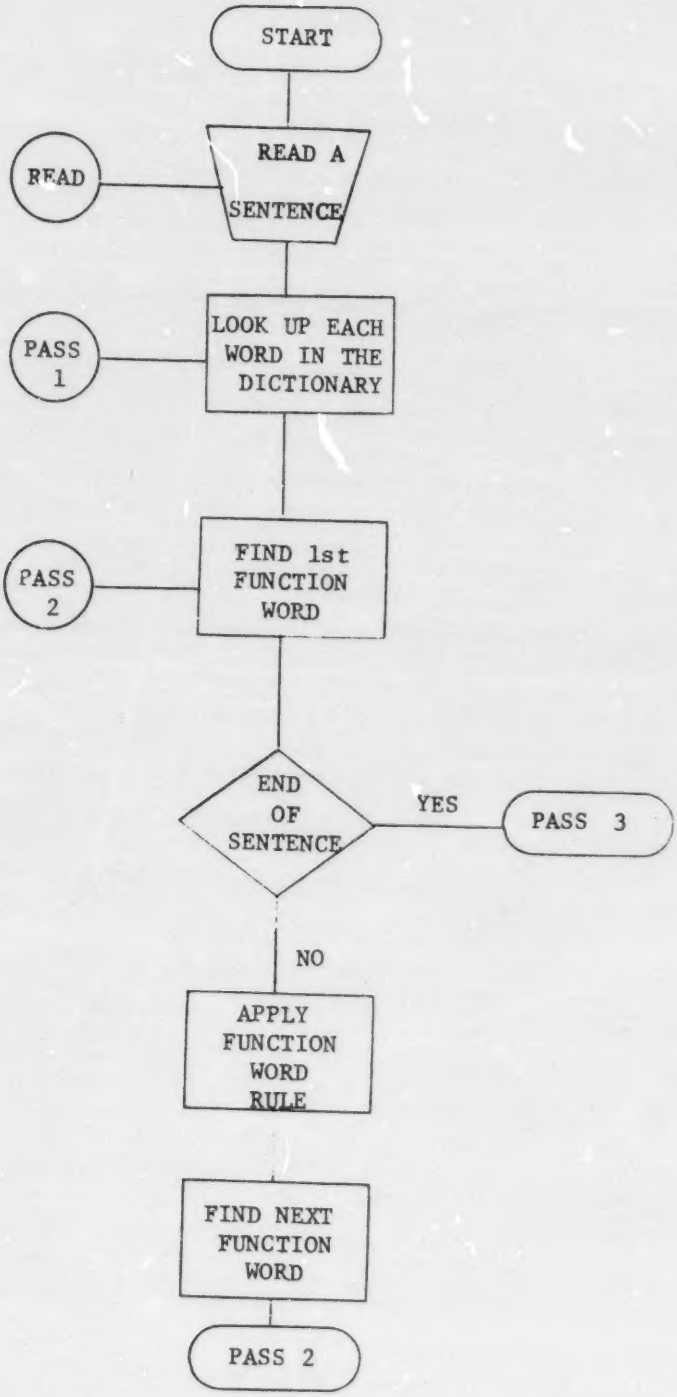


Figure 4.1 Flowchart of MYRA

Figure 4.1 (Continued)

AUXILIARY VERBS		CONJUNCTIONS	DETERMINERS	DETERMINERS AND PRONOUNS
IS	YOU'LL	AND	A	THAT
WAS	DOESN'T	BUT	HIS	THIS
BE	LET'S	OR	THE	ONE
HAD	SHE'D	NOR	AN	HER
ARE	YOU'VE		THEIR	ALL
HAVE	ONE'S		ITS	OTHER
WERE	THEY'RE		TWO	SOME
WOULD B	WE'LL		FIRST	THESE
BEEN	WE'RE		MY	ANY
HAS	WHAT'S		OUR	MANY
WILL		EXPLETIVES	MOST	EACH
CAN			YOUR	THOSE
COULD		NOW	THREE	BOTH
MAY		WELL	1	SAME
DO		OH	EVERY	ANOTHER
DID			2	FEW
MUST			FOUR	SOMETHING
SHOULD			FIVE	EITHER
BEING			3	NEITHER
MIGHT			SIX	WHATEVER
DON'T			MILLION	NONE
DOES			4	
GOT			HUNDRED	
DIDN'T			TEN	
DONE			10	
IT'S			5	
I'M			ONES	
SHALL			6	
CANNOT			15	
THAT'S			36	
I'LL			8	
COULDN'T			12	
CAN'T			7	
WASN'T			20	
YOU'RE			25	
WOULDN'T			100	
HE'S			50	
I'VE			9	
MAN'S			11	
THERE'S			BILLION	
WON'T			THIRTY	
HADN'T			FIFTEEN	
HE'D			DOZEN	
ISN'T			16	
			18	

Figure 4.2. The Dictionary for the Grammatical Class Assignment Program (MYRA).

INTENSIFIERS		PREPOSITIONS	PRONOUNS	RELATIVE PRONOUNS
MORE	LARGELY	TO	HE	WHICH
SO	ASIDE	IN	IT	WHEN
ONLY	CLOSELY	WITH	I	WHO
NOW	HIGHEST	AS	THEY	WHAT
SUCH	TOMORROW	ON	YOU	THEN
EVEN	CLOSER	AT	SHE	WHERE
ALSO	HEAVENLY	BY	WE	NOW
MUCH	SOMEHWERE	FROM	HIM	WHILE
JUST	WIDELY	OUT	THEM	WHY
TOO	GRADUALLY	UP	ME	WHOSE
VERY	REAR	ABOUT	US	WHOM
STILL	EXTREMELY	INTO	HIMSELF	
HERE		LIKE	NOTHING	
NEVER		OF	OTHERS	
AGAIN		FOR	ITSELF	SUBORDINATE
ONCE		AFTER	ANYTHING	CONJUNCTIONS
ALWAYS		BEFORE	THEMSELVES	
LESS		THROUGH	EVERYTHING	IF
ALMOST		BACK	ANYONE	THAN
ENOUGH		DOWN	MYSELF	BECAUSE
BETTER		BETWEEN	HERSELF	SINCE
LATER		UNDER	EVERYONE	HOWEVER
RATHER		AGAINST	SOME	THOUGH
OFTEN		DURING	NOBODY	YET
EARLY		WITHOUT	EVERYBODY	ALTHOUGH
ALONG		AROUND	YOURSELF	THUS
EVERY		UPON	SOMEBODY	PERHAPS
LEAST		UNTIL		WHETHER
REALLY		TOWARD		ELSE
AGO		PER		
FULL		AMONG		
FURTHER		WITHIN		
USUALLY		ABOVE		
SOON		BEHIND		
NEAR		OUTSIDE		
THIRD		EXCEPT		
GREATER		INSTEAD		
LATE		AHEAD		
HIGHER		BESIDE		
EARLIER		BESIDES		
FORMER		TOWARDS		
LOWER		ONTO		
LARGER		TILL		
YESTERDAY				
SOMEHOW				

Figure 4.2 (Continued)

ADJ	adjective	'NO'	the word 'no'
ADV	adverb	NON	noun
AUX	auxiliary verb	'NOT'	the word 'not'
CNJ	conjunction	'THERE'	the word 'there'
DTR	determiner	PNC	punctuation
EOS	end-of-sentence punctuation	PRN	pronoun
EXP	expletive	PRP	preposition
INT	intensifier	REL	relative pronoun
DR2	pronouns and determiners	SUB	subordinate conjunction
VRB	verb	PTC	participle
		GER	gerund

For logical operators, the following symbols are used:

XXX	unidentified element
---	any element
~	not
⇒	yields
|	or
...	continuation

The rules and operation of the analyzer for MYRA can best be illustrated by some simple examples. MYRA accepts as input ordinary English text punched in machine-readable form. No special coding of the text is necessary; words are delimited by one or more blanks, and the beginning and end of each word is checked for punctuation.

For each class of function words, a set of rules has been developed which accounts for the majority of the patterns in which each function word occurs. The rules which have been developed for the class of prepositions should present a good illustration of the various types of rules which MYRA embodies.

Function word rules are applied in a linear fashion as each function word is encountered. For example, suppose the sentence being processed were:

The train was on the old wooden bridge.

At the end of the first pass, MYRA would replace the sentence by the following array of elements:

DTR XXX AUX PRP DTR XXX XXX XXX EOS

Since the first unidentified element follows a determiner, it is either a noun or an adjective. Since there is only one blank element, this element must be a noun. The three blank elements remaining are recognized as a noun phrase and are assigned the classification ADJ ADJ NON. Hence at the end of the second pass, the array has the form:

DTR NON AUX PRP DTR ADJ ADJ NON EOS

If a determiner is followed by several consecutive blank elements, the last of the series is identified as a noun, and all other blanks are identified as adjectives. This is one of the patterns which is sought whenever a determiner or a preposition is encountered;

(DTR PRP) XXX ... XXX --- (DTR PRP) ADJ ... NON ---

A few examples of other rules which are applied when a preposition or a determiner is encountered is the second pass (different rules are applied during the third pass) include:

1. ... ('TO')XXX --- ... ... ('TO') VRB --- ...
2. ... (DTR PRP) INT XXX XXX ... ... (DTR PRP) INT ADJ XXX ...
3. ... (DTR PRP) XXX ... PRN ... ... (DTR PRP) ADJ ... PRN ...
4. ... PRP ... ('ED' 'ING') ... --- ... PRP ... PTC ... ---
5. ... PRP ... ('ED' 'ING') --- ... ... PRP ... GER --- ...
6. ... (DTR PRP) 'NO' ... ... (DTR PRP) ADJ ...

These examples illustrate the simplistic nature of the most of the rules which have been incorporated in MYRA and which have contributed to the programs' efficiency and effectiveness.

The third pass determines first that each element has been assigned to a grammatical class. In the extreme case, a sentence would contain no function words. MYRA would therefore assign a default pattern to the sentence. For example, in the sentence:

Ten students passed rigid exams.

MYRA would identify the middle element of the array as a verb, and the elements to the left and right would be treated as nouns. For this example, MYRA would produce the following:

ADJ NON VRB ADJ NON EOS

The rule which assigns this pattern is:

--- XXX ... XXX EOS ADJ ... NON VRB ADJ ... NON EOS

Another case which might occur is that of several consecutive blanks in an array for which no assignment has been made. For example:

Ten students were in the class.

Upon reaching the third pass, the array would have the following assignments:

--- --- VRB PRP DTR NON EOS

First, a check is made to see if a verb has already been assigned to the sentence. Since this is the case here, the following rule would be applied:

XXX ... XXX --- ... VRB ... ADJ ... NON --- ... VRB ...

If the sentence had not contained a verb, a different rule would be used:

XXX ... XXX --- ... (VRB) ...

ADJ ... NON VRB --- ... (BRB) ...

If a sentence reaches the final pass and does not contain a verb assignment, even though every element has already been assigned, MYRA attempts to reassign an element as a verb. In the sentence:

He ran well.

before the final pass, the array would be assigned,

PRN NON ADV EOS

But the analyzer would apply to the rule:

... ADJ ... ADJ NON ... (VRB) ...

... ADJ ... NON VRB ... (VRB) ...

and the array would be correctly assigned a

PRN VEB ADV EOS

In trying to assign a verb, patterns containing determiners and prepositions are again examined. Some of the rules which might be applied in this phase are:

1. DTR ADJ ... ADJ NON DTR DTR ADJ ... NON VRB DTR ...
2. DTR ... ADJ NON ('OF' 'WITH' 'IN' NON VRB ('OF' 'WITH' 'IN')
3. PRP ... ADJ NON ('IN' 'INTO' 'OF'

PRP ... NON VRB ('IN' 'INTO' 'OF') ...

This gives a sample of the function-word rules which MYRA uses to make grammatical assignments. By applying these rules as each function word is encountered, great flexibility is achieved in the types of sentences which can be processed. These rules were developed to handle

well-constructed English sentences; however, because of the flexibility which the rules provide, MYRA is also able correctly to identify sentence fragments, and other "poorly formed" sentences. This capability is important to preserving the data of a conversation, which often contains sentence fragments or other "poorly formed" utterances. The sample dialogue presented at the beginning of this document was processed through the MYRA program. Each sentence is presented below as it would appear after being processed by MYRA.

1. Dana succeeded in putting a penny in a parking meter today without

   NON VRB PRP GER DTR NON PRP DTR ADJ NON ADV PRP

   beign picked up.

   GER PART PRP

2. Did you take him to the record store?

   VRB PRN VRB PRN PRN DTR ADJ NON

3. No, t. the shoe repair shop.

   EXP PRP DTR ADJ ADJ NON

4. What for?

   RPRN PRP

5. I got some new shoe laces for my shoes.

   PRN VRB DTR ADJ ADJ NON PRP ADJ NON

6. Your loafers need new heels badly.

   ADJ NON VRB ADJ NON ADV

4.2.4 Conclusion MYRA has been tested on both technical and non-technical text totaling over 14,400 words. The current implementation of MYRA was programmed in PL/I for the IBM SYSTEM/360 Model 75 using 32,704 (8-bit) bytes. Although processing time has been difficult to measure precisely, MYRA can process at least 8300 words per minute. This is clearly a worst case because of the way in which times are reported by the computer center.

An accuracy of over 91% was achieved in these analyses. Several things should be noted about the determination of the correct grammatical class. First, all participles and gerunds were counted as correct if they were labelled adjective or noun respectively. Second, if a word

which is normally classified only as a noun, such as the word circuit, were used as an adjective as in circuit breaker this was also counted as correct.

Most of the errors were due to incorrect assignments made to words which should have been classified as adjectives, nouns, verbs, or adverbs. There were few cases in which pronouns and determiners were incorrectly identified and a few function words not included in the dictionary were also misclassified. Verbs were the most difficult class to identify and were a frequent source of error. When a verb was incorrectly identified, this often caused errors in the classification of other words. For example, consider the following sentence.

Put the top by the child sitting in the corner.

This sentence is incorrectly processed by MYRA. The word sitting is misclassified as a noun which causes the word child to be misclassified as an adjective.

Below is presented the sentence (a) as assigned by hand and (b) as assigned by MYRA:

(a) VRB DTR NON PRP DTR NON PARTICIPLE PRP DTR NON

Put the top by the child sitting in the corner.

(b) VRB DTR NON PRP DTR ADJ NON PRP DTR NON

This kind of error is the source of the majority of errors made by MYRA.

The parsing operations produced by MYRA prepare the data for phase two of CALAS, the phrase grouping procedure.

4.3 The Second Phase of CALAS: Phrase Grouping Process

After achieving a high degree of accuracy with grammatical class assignment, we then developed an algorithm to effect clause separation and to assign case roles. As we studied the output from this procedure, we noticed once again that data were being lost. Even when all nouns in a sentence were accurately assigned, much of what was put out there by the participants (e.g., single word adverbials, some prepositional phrases) was not represented in the case role array produced by these analytic procedures.

Since we were philosophically unprepared to impose value judgments on elements in the sentence, we concluded that case assignments would be more representative of the dialogue if case roles were assigned to every phrase in the clause (exclusive of the verb phrase). Whereas Fillmore (I) discusses primarily noun phrases and their relationship to the verb, we began to view every phrase within the clause as assuming a case role. Consequently, we decided that, in partitioning a clause

into phrases, every word in the clause is either (1) a substitute for a word group (i.e., phrase) or (2) part of a word group. These groups are similar to the phrase types outlined by Cook (6). A phrase is a unit composed "potentially of two or more words, which does not have the characteristics of a clause and which typically, but not always, fills slots at the clause level". (6). Phrases can be divided into the categories, exorelational, endorelational, and arelational, as illustrated in Table 4.1.

Table 4.1 Categories of Phrases

Category	Type	Abbreviation
exorelational	Prepositional Phrases	(PP)
endorelational	Coordinate Phrases Appositional Phrases	(NP_c, etc.)
arelational	Noun Phrase, Adjective Phrase, Verb Phrase, Adverb Phrase	(NP) (ADJP) (VP) (ADVP)

4.3.1. Exorelational Phrases consist of two constituents, a relator (preposition) and an argument (NP). This category of phrase is called exorelational because the relation is external in the construction. The relator fits the NP for special use in syntax as adverbial or adjectival.

1a. Adverbial prepositional phrases are the most common and the preposition fits the noun phrase for use as a locative, temporal, manner, accompaniment or other adverbial. Single adverbs may substitute for these phrases.

1b. Adjectival prepositional phrases are those in which the preposition fits the noun phrase for use as an adjectival, and the preposition OF is the most common preposition used in this fashion.

NOTE: The infinitive marker TO, found only in the context TO + Verb, or TO + BE + Verb, must be separated from the preposition set.

4.3.2. Endorelational Phrases consist of two kinds of constituents, coordinating conjunctions and nominals.

1a. Coordinate constructions consist of n ($n \geq 2$) separate words or phrases of the same word class, typically joined by n-1 coordinating conjunctions. Any of the major word classes may be joined, so that there exist coordinate noun phrases, verb phrases, adjective phrases, and adverb phrases.

The endorelational phrase functions as a simple phrase of the same word type, except for changes in number agreement. Thus Jack and Jill, NP_c, functions like NP but with plural agreement.

1b. Appositional constructions consist of two noun phrases which have the same semantic referant, and which are usually separated by commas. They also function as single noun phrases in the structure. Thus Tony, the barber, a NP_{app}, functions as a single NP in the sentence.

4.3.3. Arelational Phrases are constructions without relators and consist of a head word optionally preceded by a series of modifiers. These phrases are of 4 kinds, corresponding to the 4 major parts of speech: noun, verb, adjective, adverb.

1a. A noun phrase is an arelational phrase with a noun as the head word. The modifiers are generally determiners, quantifiers, and descriptive adjectives.

1b. A verb phrase is an arelational phrase with a verb as the head word. The modifiers are generally auxiliaries and negatives.

1c. An adjective phrase is an arelational phrase with an adjective as the head word. The only modifiers are intensifiers, such as VERY or MOST.

1d. An adverb phrase is an arelational phrase with an adverb as the head word. The only modifiers are intensifiers.

Based on this classification of phrases, an algorithm was developed and then implemented in PL/I to partition clauses into phrases. The program was subsequently modified so that it now partitions sentences (rather than clauses) into phrases. The input for the program is the output from the grammatical class assignment procedure (MYRA). The program consists of three parts (see Figure 4.3). Part one groups identical grammatical class assignments connected by the conjunction and so that these are treated as one part of speech in further processing. Part two is the main grouping routine and yields one of six phrases types (Adj. Phrase, Adv. Phrase, Conj. not joining identical elements, Noun Phrase, Prep. Phrase, and Verb Phrase). The third phase pairs identical phrase assignments connected by and, so that these are treated as one phrase for case role assignment purposes.

The object of Part I, the first conjoining process, is to join identical word classes into a functioning unit. Except for number agreement, the pairs of words joined by coordinating conjunctions will be found to function in the same way as a single word in the structure. In this process we would expect the following results:

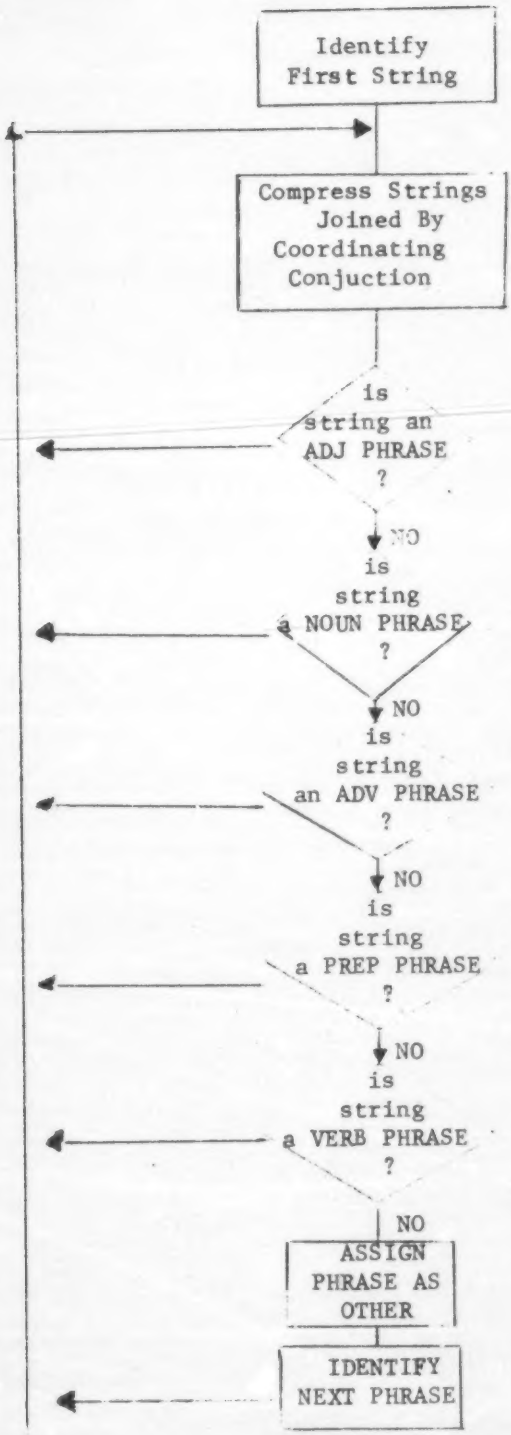


Figure 4.3 Flowchart for Phrase Grouping Algorithm

Non CONJ NON	replaced by	Non
Vrb CONJ Vrb	replaced by	Vrb
Adj CONJ Adj	replaced by	Adj
Adv CONJ Adv	replaced by	Adv
Int CONJ Int	replaced by	Int

Although the replacement is a word group or "phrase", the single word label would seem to be sufficient at this stage, as the single words take on group labels, within the main grouping routine (Part II).

The most extensive processing occurs in Part II of the program, in which the grammatical class assignment string serves as input. Each assignment string is scanned from left to right and the phrases are delimited based on the following rules:

AdjP	=	Adj	
		Int...Adj	(...=one or many)
AdvP	=	Adv	
		Int...Adv	
Conj	=	Conj (no identical pair)	
NP	=	Pn	
		N	
		Adj...N	(with N terminal)
		D...N	
		D...Adj...N	
		Int...N	
		Dtr...Int...N	
		Adj...Int...N	
		Dtr...Adj...Int...N	
PrepP	=	Prep Pn	
		Prep N	
		Prep Adj...N	
		Prep D...N	
		Prep D...Adj...N	
		Int...N	
		Dtr...Int...N	
		Adj...Int...N	
		Dtr...Adj...Int...N	
VP	=	Aux...	
		Aux Neg	(If no V found)
		Aux Neg Aux...	
		V	
		Aux...V	
		Aux Neg V	
		Aux Neg Aux...V	

Based on these definitions the assignment string is rewritten as a series of phrase labels, including:

NP VP AdjP AdvP PrepP Conj

After Part II of the program is completed, there may remain in the string conjunctions whose functions is to join identical phrases, rather than word classes. Therefore, in Part III, the CONJ connecting process is instituted again to coordinate identical phrases into a single functioning unit. For example, consider the phrase, "He and the old man". With the grammatical classes, PN CONJ DET ADJ NON, this phrase would fail the CONJ test which groups identical word classes. However, once the phrase has been labeled NP CONJ NP, it now passes the second CONJ test which groups identical phrase assignments.

Below are samples of how dialogue looks after it has been processed by both the grammatical class assignments and the phrase grouping algorithms. The first is from Hemingway's "Old Man and the Sea" and the second is the sample dialogue used throughout this report.

Clause
Numbers

```
    DTR  AUX  DTR ADJ NON
1.  He / was / an old man /
    NP   VP       NP

    RELPN   VRB      ADV     PRP DTR NON   PRP DTR ADJ NON
2.  who / fished / alone / in a skiff / in the Gulf Stream
    NP      VP       AdvP      AdvP           AdvP

    CONJ  PN  AUX VRB      ADJ NON              ADV
3.  and / he / had gone / eighty-four days / now
    Conj  NP      VP            NP             AdvP

      PRP      VRB      DTR NON
4.  without / taking / a fish
                VP       NP

    PREP DTR DTR ADJ   NON    DTR NON  AUX AUX    PREP PN
5.  In the first forty days / a boy / had been / with him.
         AdvP                  NP       VP        AdvP

    CONJ  PREP  DTR   NON     PREP  DTR  NON    DTR ADJ  NON          AUX VRB     PN
6.  But / after forty days / without a fish / the boy's parents / had told / him / that
    CONJ      AdvP              AdvP               NP                  VP       NP


    DTR ADJ NON    AUX   ADV   INT        CONJ  INT    ADJ     ADJCl
7.  the old man / was / now / definitely and finally salao / WH (S8)
    DTR ADJ NON    AUX   ADV              INT          ADJ     (by Step #1 conjoining)
        NP         VP    AdvP           AdjP                    AdjCl
```

Note: A rule would be needed here, relating the AdjP as the predicate adjective, and taking precedence over the AdvP, which can occur anywhere)

```
      PN   AUX   DTR  ADJ   NON  PREP  NON(?)
8.  which / is / the worst form / of unlucky.
     NP     VP      NP            AdjP2     (by the OF rule)
     NP     VP              NP
```

Note: Prep Phrase with the Prep OF are not part of clause structure, and are specially designated as AdjP2, meaning they must be subordinated to the preceding noun. (This rule results here in the correct structure, NP (BE) NP, with the whole phrase as the predicate nominal.) In the print-out, the OF phrase would be listed as a feature of the noun.

Sentence
Number

```
     NON        VRB       PRP  GER  DTR  NON     PRP DTR  ADJ  NON      ADV
1.  Dana / succeeded / in putting a penny / into a parking meter / today /
      NP        VP             PRP P                  PRP P          ADV P

     VRB   PRN   VRB   PRN   PRP DTR  ADJ    NON
2.   Did / you / take / him / to the record store?
           NP    VP     NP       PRP P

     EXP   PRP PTR ADJ   ADJ   NON
3.   No  / to the shoe repair shop.
     EXP      PRP P

     REL   PREP
4.   What for?
       NP

    PRN  VRB   DTR  ADJ ADJ   NON     PRP ADJ  NON
5.   I / got / some new shoe laces / for my shoes.
     NP  VP           NP                PRP P

     ADJ   NON        VRB   ADJ NON      ADV
6.   Your loafers / need / new heels / badly.
          NP          VP      NP         ADV P
```

4.4 Third Phase of CALAS - Clause Separation

After the text has been processed by the grammatical class assignment and phrase grouping programs, it is ready for clause separation procedures, the final phase of parsing, before case role assignments can be made. As mentioned earlier, a clause is a string of words with one and only one predicate (6). As Cook (14) points out, a predicate is not simply a verb form, it is a "verb form filling the predicate slot" in a structure. Each verb form used as a predicate constitutes a clause, and in a text there can be no fewer clauses than there are verb forms. According to his classification there are several types of clauses and each of these is partitioned in the automated clause separation process: (1) independent clauses, (2) subordinated clauses, (3) relative clauses and (4) partial clauses which contain, as predicate, non-finite forms of the verb such as

infinitives and participles. Clause types 2, 3, and 4 are all dependent clauses. Algorithms have been written and implemented in PL/I which separate and label each of these clause types.

4.4.1 Independent Clause Separation. Compound sentences are those in which there is more than one independent clause in parallel relationship. Such clauses usually contain a coordinating conjunction. When one is not present, the coordinating conjunction slot will probably be filled by a semicolon. Each of these independent clauses may have one or more dependent clauses but the dependent clauses are separated at a later stage in the processing. Only the independent clauses are separated at this stage.

The separation of independent clause candidates occurs as follows:

1) single coordinator or punctuation mark

fragmentation points such as ; AND, OR, BUT are located and the word string is separated before the connector or coordinator*

2) double connectors

when the word string contains double connectors, the connectors or coordinators appear in pairs one each in front of both independent clauses. Examples of double connectors include: Both...And, Either...Or, Neither...Nor. The string is separated immediately before the second connector.

Below is an example of how the clause separation looks at this point in the processing:

Clause
Number

1. He was an old man who fished alone in a skiff in the Gulf Stream.

2. AND he had gone eighty-four days now without taking a fish.

3. In the first forty days a boy had been with him.

*The reader may recall that compound word entries and compound phrase entries have already been joined in the phrase grouping process. These then are "out of the way", not subject to further scanning; therefore, are not inadvertently chosen as independent clause fragmentation points.

4. BUT after forty days without a fish the boy's parents had told him that the old man was not definitely and* finally salao which is the worst form of unlucky.

5. AND the boy had gone at their orders in another boat which caught three good fish the first week.

6. It made the boy sad to see the old man come in each day with his skiff empty.

7. AND he always went down to help him carry either** the coiled lines or** the gaff and *** the harpoon and *** the sail that was furled around the mast.

8. The sail was patched with flour sacks.

9. And furled, it looked like the flag of permanent defeat.

The clause separation process now continues with the partitioning of subordinated clauses.

4.4.2 Dependent Clause Partitioning: I. Subordinated Clause Separation
Dependent clauses can be categorized based on their form as subordinated, relative, or partial. A subordinated clause contains a relator and an independent clause; a relative clause contains a relative pronoun; a partial clause, a predicate that is either an infinitive or a participle.

A subordinated clause consists of an independent clause that is introduced by a subordinating conjunction (SUB-CONJ or SC). The combination of the subordinating conjunction and independent clause acts as a dependent clause. The algorithm for separating subordinated clauses identifies a subordinating conjunction, and divides the word string immediately before it. Different conjunctions signify different functions for a dependent clause, but whatever the lexical entry of a subordinating conjunction the clause structure and rules for fragmentation remain the same.

For example:

(a) When I come, he will go.

(b) If I come, he will go.

Each example contains an introductory subordinated clause followed by an independent clause but example (a) signals time and (b) signals condition.

---

*Note by the conjoining rules discussed in Section 4.2, Int + Conj + Int becomes Int, so the AND between definitely and finally is not a clause fragmentation point.
**The phrase conjoining rules eliminate these conjunctions from incorrect clause separation processing.
***The word conjoining rules eliminate these conjunctions from processing.

The algorithm for subordinated clause separation follows.

1. Input word string with grammatical class assignments and phrase grouping markers.

2. Scan for SUB CONJ*

3. If found, scan for VRBP before and after

4. If found, cut string before SUB CONJ

5. If not found (probably means the subordinate clause is in introductory position), scan for 2 or more VRBP after SUBJ CONJ but before Coordinating CNJ marker

6. **If found, scan internal structure for SUB CNJ NP VP...punctuation...NP VRB

7. Cut string after the punctuation (e.g. SUB CNJ...NP...VP/.. ...NP...VRVP...)

   If no punctuation found, use following fragmentation points

   a) SUB CNJ NP...VRBP...NP/NP...VRBP...
   b) SUB CNJ NP...VRBP.../NP...VRBP
   c) SUB CNJ NP...VRBP /VRBP...
   d) SUB CNJ VRBP (part) /VRBP...

In addition to the subordinated clauses above, there are other subordinated nominal clauses which generally have a zero connector or are introduced by THAT. The THAT referred to in this context is not the THAT used as a relative pronoun but rather the THAT used as a detachable subordinating conjunction.

For example:

(a) He knew he was right.

(b) He knew that he was right.

When there is a zero connector the fragmentation points are:

...NP...VRBP.../NP...VRBP...

* SUB CONJ are considered function words and so are the MYRA dictionary terms listed in Figure 4.2. A partial list of subordinating conjunctions includes if, although, where, because, and unless.
**.... signifies any phrase but SUB CONJ, NP OR VRBP.

For a detachable THAT construction the fragmentation points are:

...NP...VRBP...NP.../THAT...NP...VRBP...
...NP...VRBP.../THAT...NP...VRB...

4.4.3 Dependent Clause Partitioning: II. Relative Clause Separation. A relative clause is a dependent clause that has no detachable relator. The signal for these clauses and the sign of their dependency is a relative pronoun* usually found in the first position in the clause. A relative pronoun serves both as a clause constituent and a relator and its presence signifies dependency. Compare the structure of the subordinated clause in example (a) with the relative clause in example (b).

(a) SUB CONJ
Since I left, many changes have taken place.

(b) Relator + Prn
I know what is important.

Relative clauses can be adjectival, nominal, or adverbial depending upon their position in the string. Relative pronouns can serve as subjects, objects and adverbs and are often used in indirect questions. The fragmentation rules for separating relative clauses follow the same pattern as those for separating subordinated clauses. If the relative clause is in the final position, it is isolated by using the relative pronoun as the fragmentation point. But if the relative clause is in the initial position or in the middle of the string in an adverbial or adjectival position, the internal structure of the clause will have to be scanned to find the NP VRBP combination that make up the independent clause and the RPN VRBP combination that make up the dependent clause. As in the subordinated clause, punctuation can serve as a fragmentation point. The algorithm for separating the relative clauses follows:

(1) Input word string with grammatical class assignments and phrase grouping markers

(2) Scan for relative pronoun

(3) If found, scan for VRBP before and after

(4) If VRBP found, cut string before RPN

(5) If not found, scan for 2 or more VRBP after the RPN but before coordinating conjunction or end of sentence marker

*The relative pronouns are listed in the MYRA dictionary (as in Figure 4.2) and include such words as who what when where how.

(6) If found, scan internal structure of clause for ...NP, RPN...VRBP...,...VRBP...*

(7) If found set off string between punctuation as relative clause, and put remainder of string together from left to right as independent clause

(8) If no punctuation found, the following fragmentation points are used

*RPN...VRBP...NP/...NP...VP...
NP.../RPN...VRBP...NP/...VRBP...
NP.../RPN...VRBP.../VRBP...

The data are now ready to be separated by the last phase of the clause partitioning process, partial clause separation.

4.4.4 Dependent Clause Partitioning: III. Partial Clause Separation. Another form of dependent clauses is the partial clause. Partial clauses are those which non-finite forms of the verb, such as infinitives and participles are used as the predicate. A participle is an ing or ed form of the verb. It fills a predicate slot except when it is found in a modifier position. In those cases participles are treated as adjectives. For example: "the slow lumbering covered wagon" is not a clause. Participial clauses are found in nominal, adjectival or adverbial constructions. For example:

without taking a fish - nominal

the girl, eating the sundae, - adjectival

the girl sat eating a sundae - adverbial.

The fragmentation rules for separating clauses that contain participles as predicates are based on the assumption that participles do not have subjects.** Participial clauses are fragmented by identifying the participle in the word string and cutting directly before it, except when a possessive pronoun appears before the participle. In this case the fragmentation point occurs before the possessive. For instance

(a) I remember/sitting alone in the room.

(b) I remember/my sitting alone in the room.

---

*The NP, RPN VRB, VRB are essential for this fragmentation; the other elements (indicated by...) are optional.

** Cook (14) believes and we agree that a case can be made for asserting that participles do not take subjects. For example, in his extensive work analyzing Hemingway's The Old Man and the Sea, Cook discusses the following example:

"I remember you throwing me into the bow of the boat."

Cook classifies the underlined word string as a participial adjectival clause modifying you. He maintains if the sentence had read "I remember you throwing me into the bow of the boat", the string would be classified as a nominal participial clause object of the verb "to remember".

The fragmentation rules for these two cases are:

...NP...VRBP.../PARTICIPLE...

...NP,,,VRBP/POSS PRN PARTICIPLE...

If a partial clause containing the participle is not in the last position in the string, then the internal clause structure and punctuation must be scanned to determine the fragmentation points. These scanning procedures are similar to those used to separate subordinated and relative clauses.

The algorithm for separating partial clause containing participles follows:

1) Input string with grammatical class assignments and phrase grouping markers.

2) Scan for VRBP and PARTICIPLE (PRT)
   If VRBP appears before participle in the string use following fragmentation points

   ...NP...VRBP.../PRT...

   ...NP...VRBP.../POSS PRN PART...

3) If VRBP does not appear before PRT, check for punctuation.
   If found, use following fragmentation points

   NP..../...PRT..../...VRBP... (The clause set off by commas is the dependent clause. The independent clause is formed by re-writing the remaining elements in a string from left to right.)

4) If no punctuation is found, use the following fragmentation points

   PRT.../VRBP...

   PRT...NP.../NP...VRBP...

   PRT NP/NP...VRBP

   PRT...NP/...VRBP

   NP.../PRT...NP/VRBP...

Another type of partial clause is the infinitive clause. These clauses have the infinitive form of the verb as the predicate. Unlike participial clauses, infinitive clauses can contain subjects. The subject of an infinitive often performs two different functions in the sentence: one in the independent clause and one in the dependent clause. For example:

"I asked him to go home."

Him fills an indirect object slot in the first clause and a subject slot in the second. While infinitives can be found in sentences as adjectival complements and nominals their primary use is adverbial, as an adverb of purpose. Whenever to can be replaced by in order to, the infinitive is used as an adverb of purpose. When the infinitive is clearly marked by to plus the base form of the verb, the procedures for separating infinitive clauses are the same as those for separating the participial clauses. As with the other dependent clauses, infinitive clauses found in the last position require less processing since the string is cut before the infinitive marker to. The algorithm for separating partial clause structures containing marked infinitives follows:

1) Input string with grammatical class assignments and phrase grouping markers.

2) Scan for VRBP and marked infinitive (MINF)

3) If VRBP appears before MINF in string cut before the infinitive marker. For example:

   ...NP...VRBP.../MINF...

   ...NP...VRBP...NP.../MINF

   ...NP,,,VRBP...NP/MINF

4) If VRBP does not appear before MINF, check for punctuation.

5) If punctuation found, use following fragmentation points.

   NP...,/MINF...,/...VRBP...

   NP...,/MINF NP,/...VRBP (The clause set off by commas is the dependent clause. The independent clause is formed by rewriting the remaining elements in a string from left to right.)

6) If no punctuation is found use the following fragmentation points

   MINF.../VRBP...

   MINF...PN.../NP...VRBP...

   MINF...NP/NP...VRBP

   NP...VRBP...NP/MINF...

   NP/MINF NP/VRBP...

   NP/MINF/VRBP

Unmarked infinitives still present us with a problem for clause separation.

For example:

(a) He heard the lion roar.

(b) He heard him roar.

In example (a) the lion roar receives the following grammatical class assignments: DTR ADJ NON. The phrase program treats these assignments as a noun phrase and so the string never becomes a candidate for clause separation. In example (b) however, roar is classified as a verb since him is correctly classified as a pronoun; and therefore in the final phase of the separation process roar is partitioned as a clause.

The final step of the clause separation process is designed to identify unmarked infinitives. If any verb phrase in the string has been included in a clause, this step cuts the string before the VRBP. To this VRBP are appended those succeeding words that are not already included in a clause. The fragmentation rule is

...VRBP.../

At the end of the clause separation process the automatic phrasing procedures are complete and the data are now ready for case role assignment.

4.4.5 Summary for Clause Separation Process Four types of clauses are separated through the automatic clause separating process. The clauses are (1) independent clauses, (2) subordinated clauses, (3) relative clauses and (4) partial clauses. The clauses are separated in that order, and the algorithm for separation is:

(1) Subroutine for separation of independent clauses Input string with grammatical class assignment and phrase grouping markers. Scan for coordinating conjunction (CCN). If found, break string before CCN. Print one clause to a line; begin succeeding clauses with CCN.

(2) Subroutine for separation of subordinated clauses Input string with grammatical class assignment and phrase grouping markers. Scan for subordinate conjunction (SCN). If subordinated clause is in final position, use SCN as fragmentation point cutting the string before it. If subordinated clause is in initial or other position, scan for punctuation (usually commas) as fragmentation point. If no punctuation found, use the following:

(a) SCN NP...VRBP...NP/NP...VRBP

(b) SCN NP...VRBP.../NP...VRBP

(c) SCN NP...VRBP/VRBP...

(d) SCN VRBP (part)/VRBP...

The fragmentation point for a zero connector (omitted conjunction) or a detachable <u>that</u> construction are:

(e) Zero connector -

...NP...VRBP.../NP...VRBP...

(f) detachable THAT-

...NP...VRBP...NPK../THAT...NP...RBP...

...NP...VRBP.../THAT...NP...VRBP

(3) <u>Subroutine for separation of relative clauses</u>

(a) Input word string with grammatical class assignment and phrase grouping markers.

(b) Scan for relative pronoun

(c) If found, scan for VRBP before and after. If found, cut string before RPN.

(d) If not found, scan for 2 or more VRBP after the RPN but before coordinating conjunction or end of sentence marker.

(e) If found, scan internal structure of clause for

...NP, RPN...VRBP...,...VRBP...

(f) If found, set string between punctuation as relative caluse and put remainder of string together from left to right as independent clause.

(g) If no punctuation found, the following fragmentation points are used.

RPN...VRBP...NP/...NP...VP...

NP.../RPN...VRBP...NP/...VRBP...

NP.../RPN...VRBP.../VRBP...

(h) The fragmentation points for deletable THAT clauses are:

...NP...VRBP...NP.../THAT...NP...VRBP...

...NP...VRBP.../THAT...NP...VRBP...

(i) The fragmentation point when the relative pronoun is deleted is:

...NP...VRBP.../NP...VRBP.

(4) Subroutine for separation of partial clauses containing participles.

...NP...VRBP.../PRT...

...NP...VRBP.../POSS PRN PART...

NP...,/...PRT...,/...VRBP...

PRT.../VRBP...

PRT...NP.../NP...VRBP...

PRT NP/NP...VRBP

PRT...NP/...VRBP

NP.../PRT...NP/...VRBP...

(5) Subroutine for separation of partial clauses containing infinitives.

...NP...VRBP.../MINF...

...NP...VRBP...NP.../MINF

...NP...VRBP...NP/MINF

NP...,/MINF...,/...VRBP...

NP...,/MINF/NP,/...VRBP

MINF.../VRBP...

MINF.../VRBP...

MINF...NP.../NP...VRBP...

MINF NP/NP...VRBP

MINF...NP/...VRBP

NP...VRBP...NP/MINF...

NP/MINF NP/VRBP...

NP/MINF/VRBP

(6) Final subroutine for clause separation.

Scan any remaining words in the string for VRBP if found, fragment as follows:

.../VRBP.../

4.5 The Fourth Phase of CALAS: Computer assisted Case Role Assignment

After the text has been processed by the grammatical class assignment, phrase grouping and clause separation algorithms, it is then ready for case role assignment. An algorithm for case role assignments has been written and implemented in PL/I.

Recall that case grammar proposes that grammatical structure consists of a series of non-linearly ordered case marked noun phrases associated with a verb phrase. Cases may be viewed as roles which retain their character while participating in different natural language utterances.

The case-role-assignment program consists of three parts (see Figure 4.4). Part I classifies the verb; Part II assigns essential cases; Part III assigns peripheral cases.

4.5.1 Verb Classification The pivotal word class in analysis via case grammar is the verb, therefore the verb governs the essential cases that may surround it. A mistake in verb classification results in incorrect case role assignments. The verb types delineated by the case-role-assignment program are State, Benefactive, Experiencer, and Agentive. (The reader will recall that these verb types are defined in Figure 3.1.) Figure 4.5 presents a more extensive classification used by some linguists who make case role assignments by hand (23). There are several differences between the verb classification procedures used by CALAS and those presented in this figure. For example, Agentive verbs are further divided into Action and Action Process verbs depending upon whether they are used transitively or intransitively in the clause. But the major difference between the CALAS classification and that of Figure 4.5 is an additional classification, Process verbs. To date we have treated these as agentive verbs. Since process verbs must be designated by dictionary entries, we are reluctant to include them as a separate type for automatic analysis. The increase in accuracy that has been gained from additional dictionary entries that we have introduced on a trial basis has not compensated for the additional processing time. Also a particular lexical item might be an agentive verb in some instances and a process in others. For example, compare the following sentences:

```
          A                    O
(a)  Jim / broke / the baseball bat.

          O                L
(b)  The key / broke / in the lock.
```

In the sentence (a) broke is considered an agentive verb as Jim instigated the breaking action. In sentence (b) broke is considered a process verb as the predication is explaining what happened to the key. Jim is assigned an agent case and key and baseball bat are both assigned object cases (both key and bat underwent some process). We are considering means other than dictionary additions to expand verb classification (see Section 6.1).

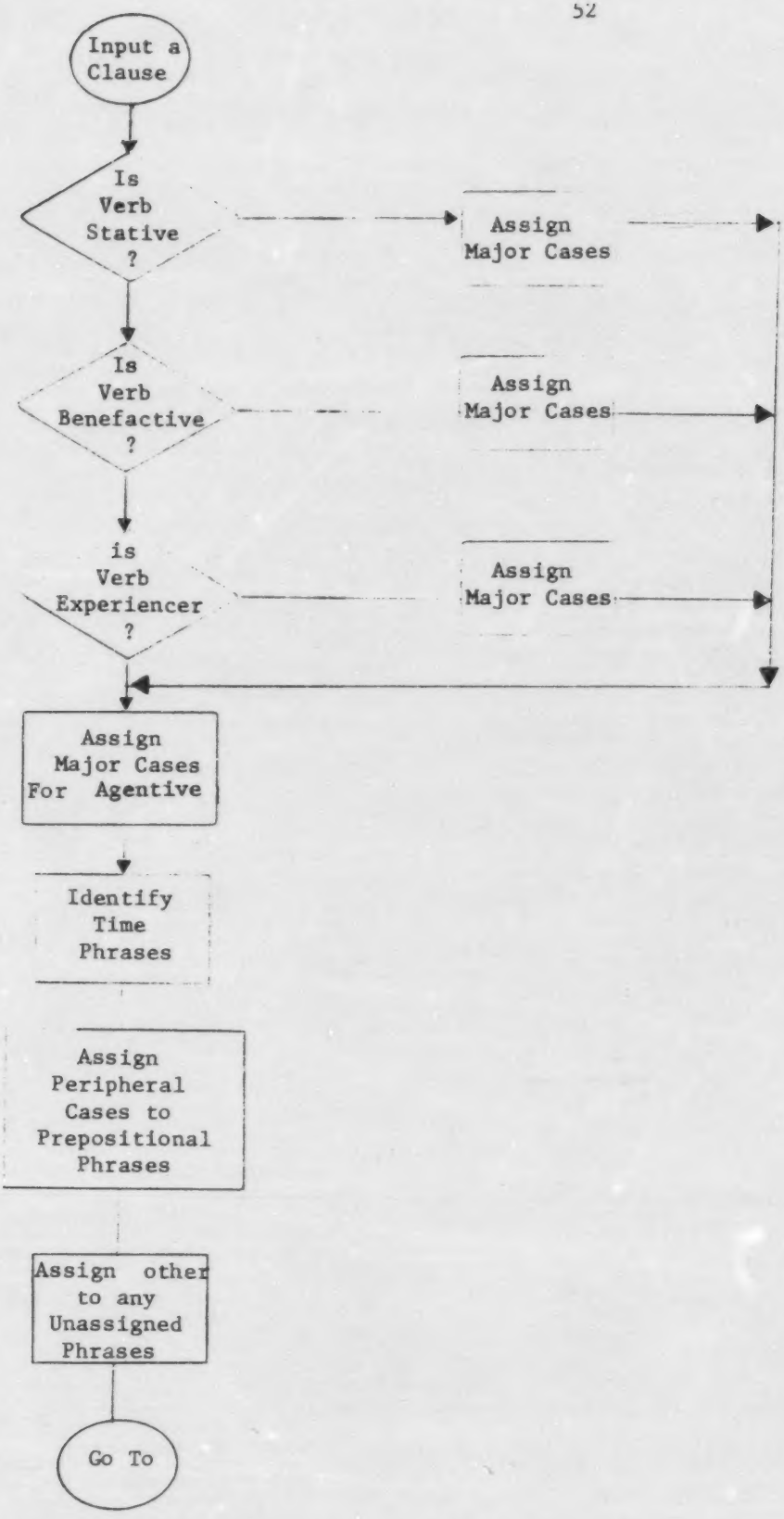


Figure 4.4 Flowchart for Case Assignment Program

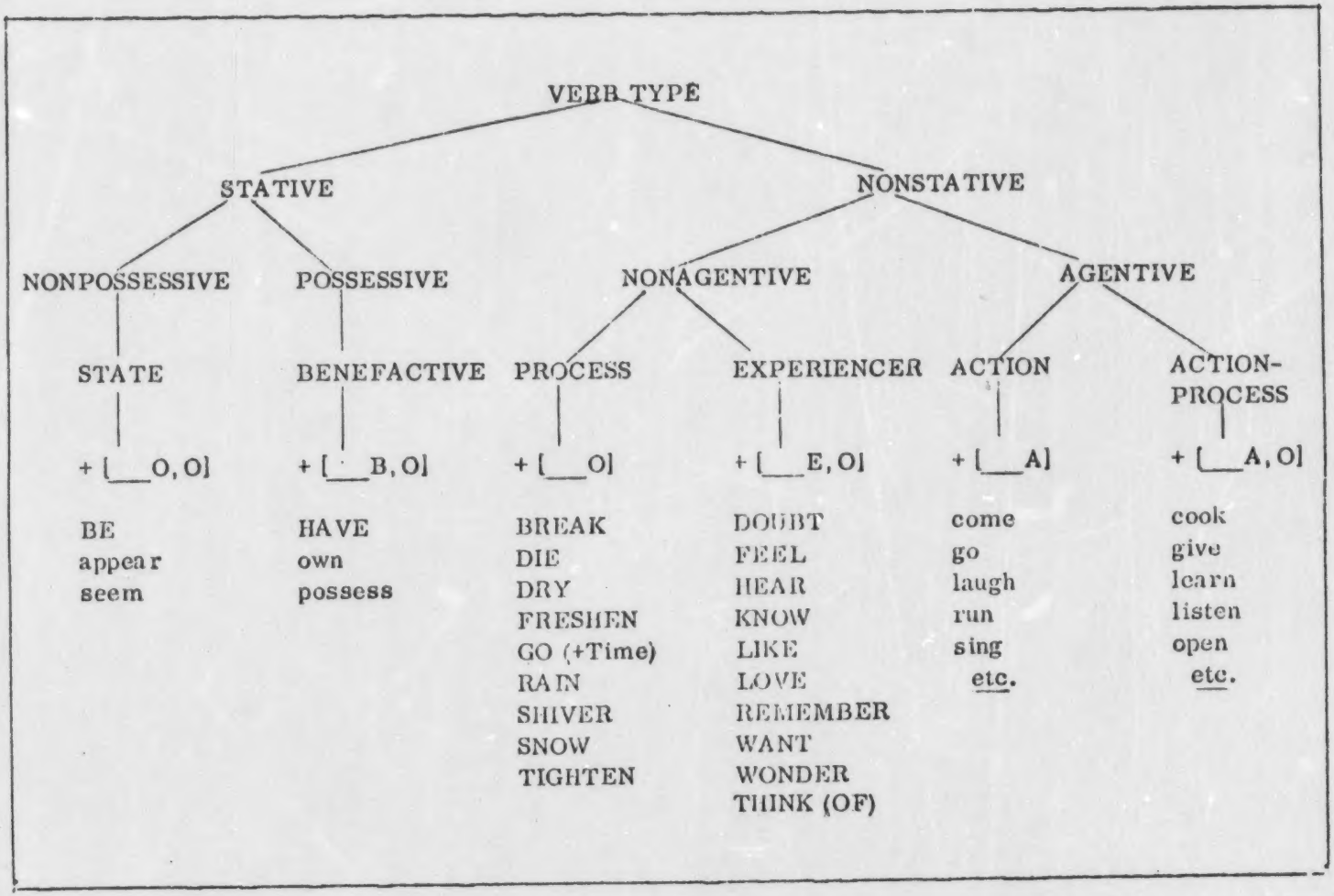


Figure 4.5 Verb Types and Case Roles.

4.5.2 Essential Case Assignment Part II of the program involves the assignment of essential cases. Essential cases are defined by the kind of verb found in the clause. These cases are also called nuclear or propositional cases. (A definition of each case is found in Section 3, Figure 3.2) After the verb phrase has been classified as State, Benefactive, Experiencer or Agentive, case assignments are made on the folling basis:

State	(O O)*
Benefactive	(B O)*
Experiencer	(E O)*
Agentive	(A E O)**

4.5.3 Peripheral Case Assignments Part III of the program assigns the peripheral or model cases. These cases occur frequently with a wide variety of utterances, such as prepositional phrases and single word adverbials, but are not governed by the type of verb contained in the utterance. Figure 4.6 summarizes the case role assignments that are derived from prepositional phrases and single word adverbials. As noted in Figure 4.6, prepositions can sometimes govern essential cases as well as peripheral ones. The algorithm for peripheral case assignments involving prepositional phrases follows.

(1) A prepositional phrase introduced by the preposition of is treated as an adjectival phrase, and is attached to a preceding noun or prepositional phrase and carries its case designation

A O

(e.g., /The Pirates of Pittsburgh / beat / the Orioles of Baltimore)

(2) The preposition to (if not part of the verb phrase) takes the locative case if its object is inanimate, and the experiencer case if its object is animate. For example,

VP E

/ Give / the money / to him.

A VP L

/ He / went / to the store.

(3) The preposition by, if it appears in a passive voice construction and its object is animate, takes the agent case. If by appears in an active voice construction where its object is inanimate, it takes the manner case. For example,

*All noun phrases before the verb receive the first case role assignment designated in the parentheses; all noun phrases after the verb, the second.
**With agentive verbs all noun phrases before the verb are assigned the agentive case; all animate noun phrases after the verb, the experiencer case; and all inanimate noun phrases after the verb, the object case.

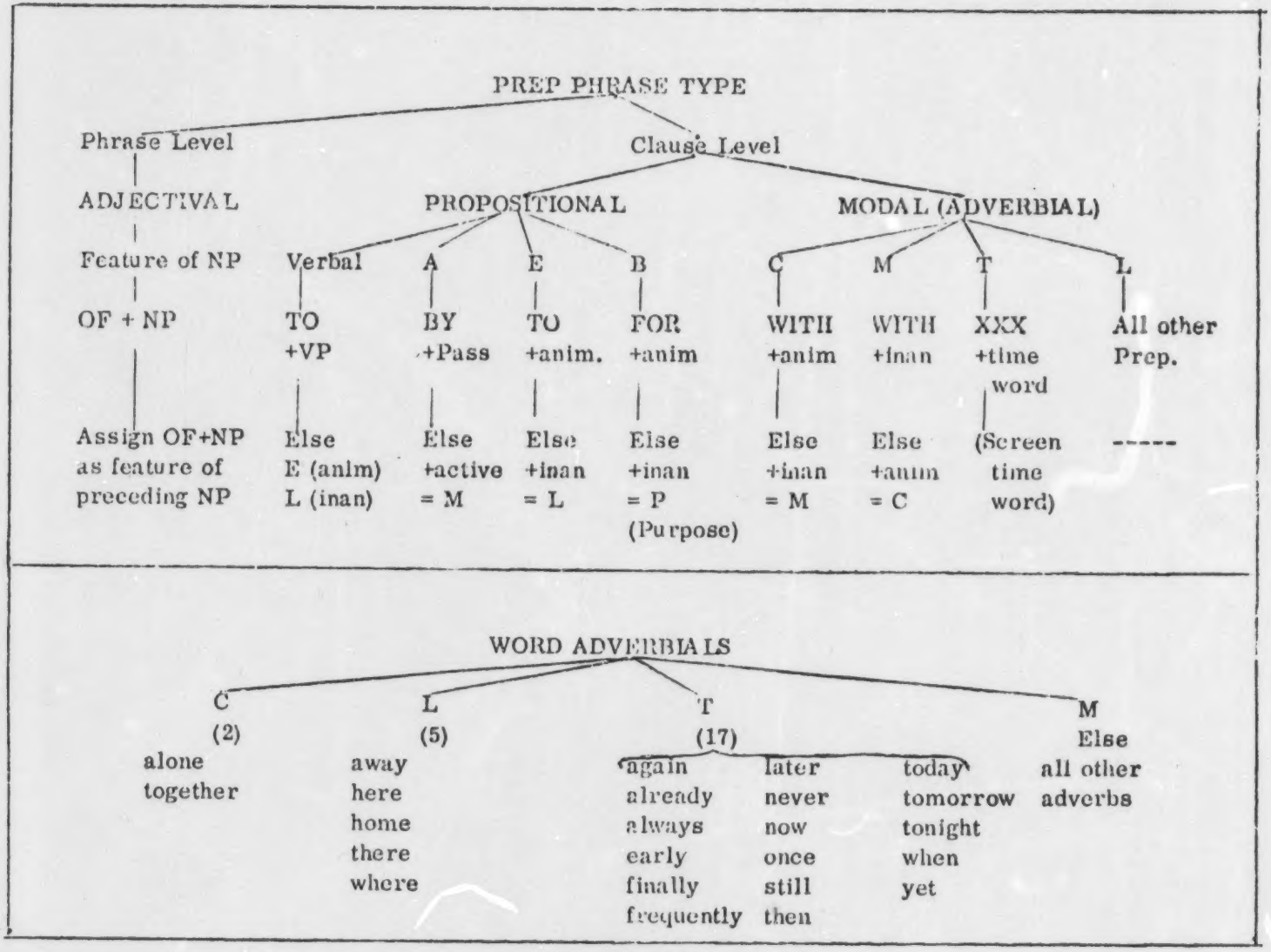


Figure 4.6 Prepositions, Adverbials and Peripheral Case Roles

```
      O          VP       A
/ The game / was won / by him.

  A    VP       O        M
/ He / won / the game / by luck.
```

(4) The preposition for takes the benefactive case if its object is animate, and the purpose case, if its object is inanimate. For example,

```
 A    VP    O     B
He / did / it / for her.
 A    VP    O     P
He / did / it / for money.
```

(5) The prepositions with and without take the comitative case if their objects are animate, and the manner case if inanimate. On, in, and from are locative. For example,

```
 A    VP       O          L
He / saw / the money / on the floor.
 A    VP                 L
He / was caught / in the rain.
 A    VP       O          L
He / took / the keys / from the table.
```

One of the most difficult peripheral cases to assign by automated means is the Time case. There seems to be no unique structural cues as to when this case occurs. Lexical entry is presently the means we use for assigning this case. Consequently, the case program has a small dictionary of time words such as time, now, day, week, minute. Any phrase except the VP in which these time words appear is assigned the T case. Similarly, (see Figure 4.6) lexical entries are the basis for determining the comitative, locative, and time cases for single word adverbials. Single word adverbials that are not included in these dictionaries are assigned the manner case. Some examples of such assignments follow.

```
 A       VP       T
He / arrived / in time.

 A       VP       T
He / arrived / yesterday.

 A                C
He / arrived / alone.

 A                L
He / arrived / home.
```

If alone were not in the dictionary, it would receive the manner case, a case role assignment that in our view is not unacceptable. Figure 3.3, Section 3, presents the case assignment for the sample dialogue.

Case role assignment, the fourth phase of CALAS, establishes the function of every phrase in the language string by assigning each a case role. A user who has processed natural language through CALAS,

then can employ automated procedures to sort the case role assignments based on criteria useful for his purposes. There sorts can be used to help him infer what is going on in language exchange. Some examples of how this might be done and some signals characterized by case role that might be important to such inferences are discussed in the next section.

5. CALAS as a Basis for Interpreting Natural Language

5.1 Information

The implementation of the four distinct phases of CALAS provide observers with a characterization of language that can be used to interpret dialogues. The advantages of CALAS are its accuracy, speed, reliability, generalizability to many communication topics and technical disciplines and its potential for being adapted for use with a number of languages. Since it is based on language used by participants, CALAS eliminates pre-processing and many individual judgments.

One way to interpret dialogue that has been processed by CALAS is to consider each case role and each verb type as a potential unit of information*. Each of these units has a somewhat different information potential, and their combinations provide different data bases from which inferences can be made. The user has a number of options available to him in deciding which potential information units to isolate from the processed dialogue. Those he chooses will depend upon his purposes for analyzing the dialogue and will, in turn, influence the inferences he makes. Rather than confining himself to one or two types of information unit, the user may extract a large number of potential information units from the dialogue and make his inferences from a very large data base. This procedure is recommended for those who wish to establish which information units have predictive value for particular situations.

5.2 Potential Information Units

Several elements from the dialogue that have the potential of becoming information for the user include: (1) use of each verb type (2) use of essential cases with each verb type (3) use of peripheral cases with each verb type (4) comparison of each speaker's use of the first three potential information units (5) comparison of use of (1), (2), and (3) across specified lexical entries or topics and (6) the sequence of the use of (1), (2), and (3) as the dialogue progresses. Figure 5.1 presents a chart for displaying these potential information units. This basic chart can be used to display specific facets of the dialogue, to compare speakers, to highlight specific topics or to illustrate any of the potential information units. As potential information units are discussed, several will be illustrated in the manner outlined in Figure 5.1 using data from the sample dialogue presented in Table 2.1. The circle in Figure 5.1 is used to represent the

*Information is defined as data of value in decision making (5).

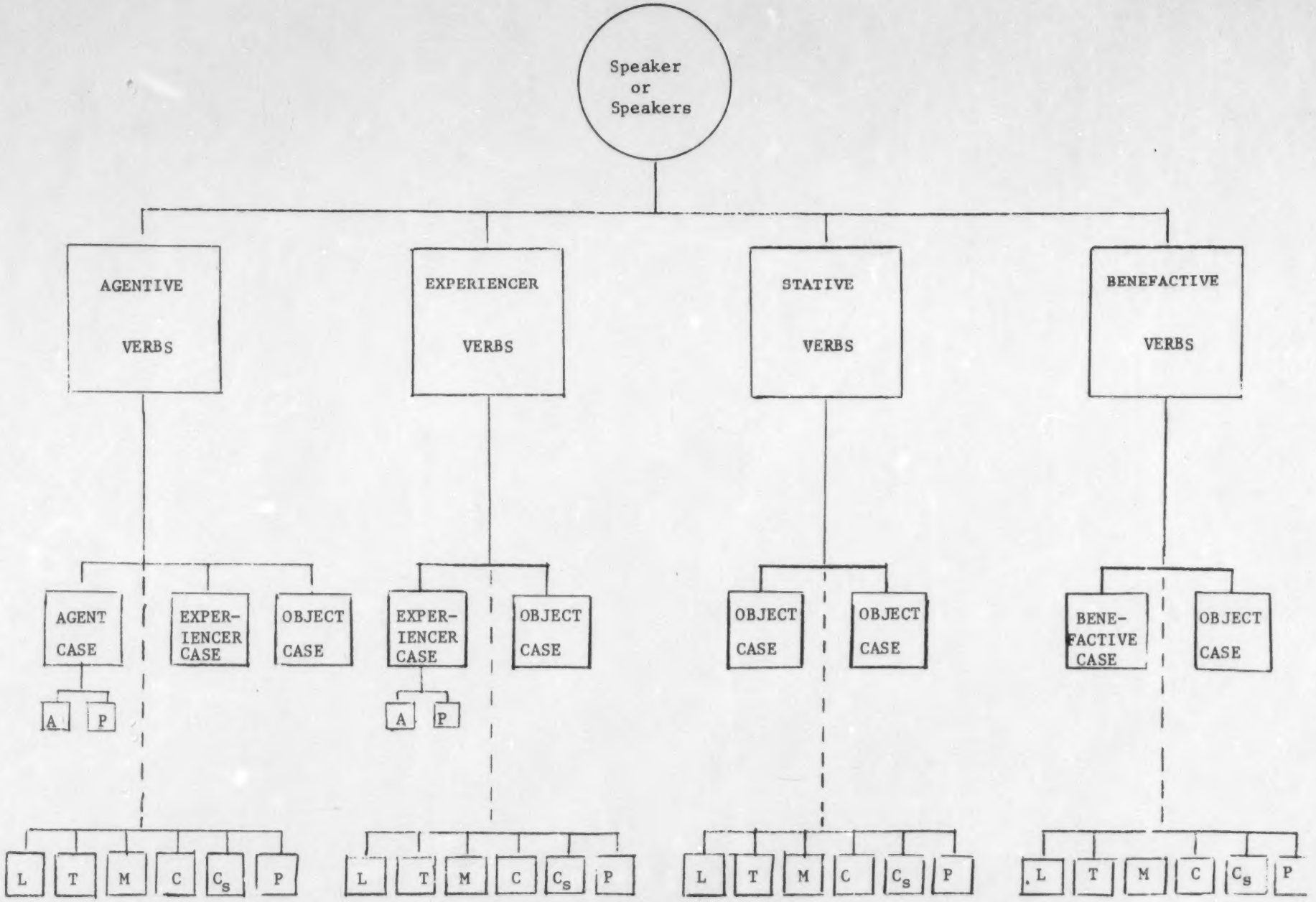


Figure 5.1 Potential Information Units
From Natural Language

speaker or speakers whose speech is being illustrated. (One could use the circle to represent a sentence or a clause.) The four large blocks represent each of the four verb types recognized by CALAS: agentive, experiencer, stative, and benefactive. The nine middle-sized blocks represent the essential cases that may appear with each of the verb types.

1. Agentive Verb: Agent Case, Experiencer Case, Object Case

2. Experiencer Verb: Experiencer Case, Object Case

3. Stative Verb: Object Case, Object Case

4. Benefactive Verb: Benefactive Verb, Object Case

In the case of the agentive and experiencer verbs there are additional slots for recording whether the case is found in an active voice or passive voice construction. The peripheral cases that may appear optionally with each of the verb types are indicated by the squares at the bottom of the chart. The abbreviations for each are as follows:

L = locative case
T = time case
M = manner case
C = Comitative
C_s = Causative case
P = Purpose case

5.2.1 Verb Types As noted earlier in Section 3 the verb type in each clause determines the essential case roles for that clause. While each essential case may not appear each time a specific verb type is used, no other essential cases may appear. The change of verb type indicates the nature of the activities being described by the speaker. For example, a speaker who is using experiencer verbs is describing feelings and possibly is discussing his relationships with others, while a speaker who is using agentive verbs is describing actions and events and if he is referring to himself at all he is doing so in a rather impersonal fashion. One might speculate that communication difficulties and other misunderstandings could easily occur between speakers if one is constantly using agentive verbs and the other, experiencer verbs. The verb type may also be indicative of how the speakers regard the topic under consideration. For example, both speakers could change to the same verb type at the time they change to a new topic. Figure 5.2 illustrates the progression of verb types found in the sample dialogue.

5.2.2 Use of Essential Cases with each Verb Type Since each verb type takes a different set of essential cases, and since each essential case need not appear each time a particular verb type appears, the use each speaker makes of essential case slots provides additional data for other participants as well as observers to a conversation. Each case role is a potential unit of information, and when the case role slots are vacant, one can study the patterns of such vacancies to make inferences about speakers' intentions and to predict dialogue outcomes. For example, omission of essential cases might indicate that a speaker

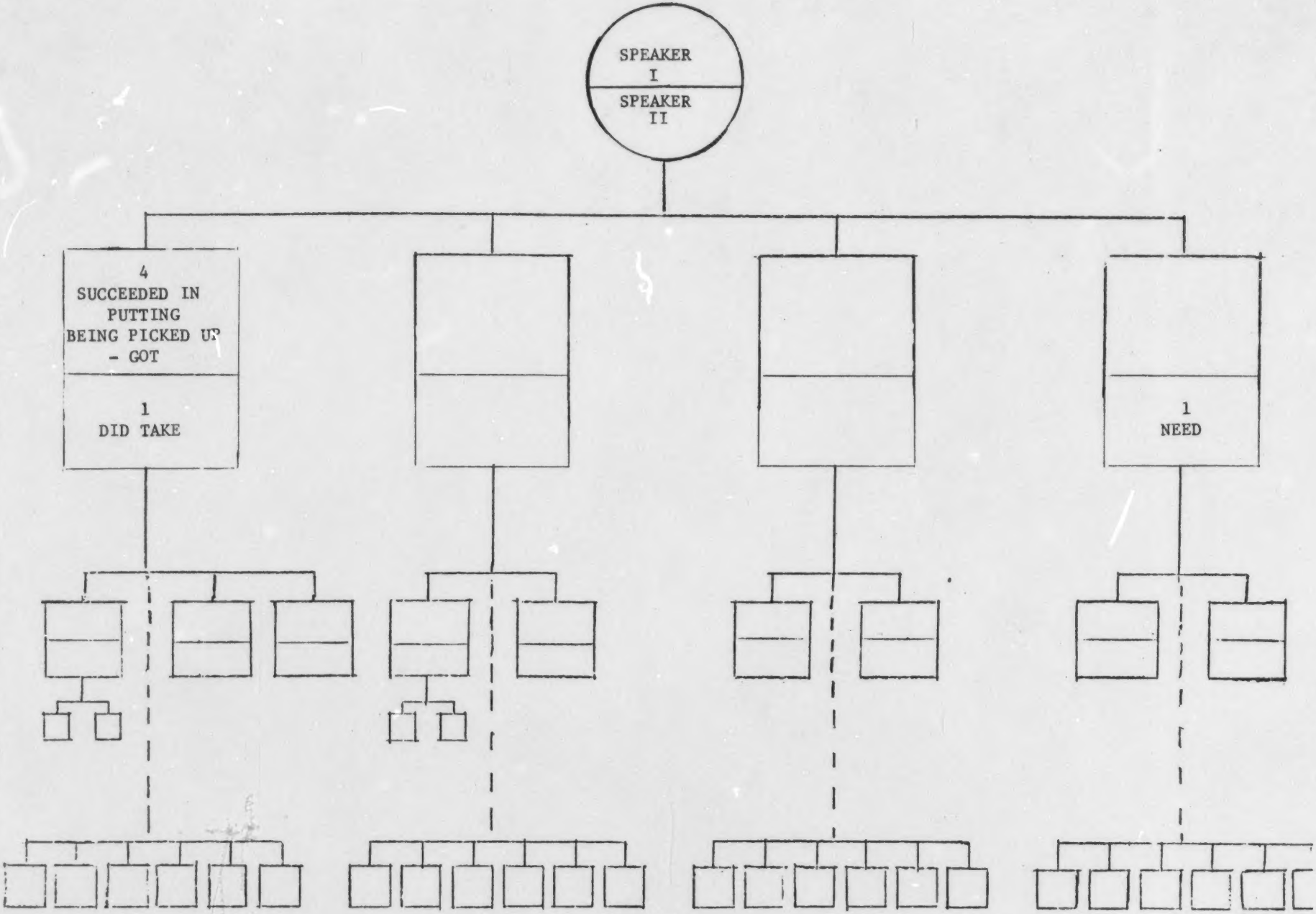


Figure 5.2 Verb Types from Sample Dialogue

(Arabic numerals indicate the number of times the speaker used the verb type indicated.)

(a) is concealing data from other participants; (b) believes that the other speakers "know" what belongs in the missing slots; (c) does not have data for those slots; (d) wishes to report only certain aspects of an activity (for stylistic purposes or other reasons, e.g., emphasis) or (e) does not need a complete set of cases to express his ideas.

Over the course of a long dialogue or text, case patterns can be identified and used as data for inferences. Knowledge of speech and writing patterns can be increased by studying case usage patterns. Such study can prove to be a good learning device for those who wish to improve their communication skills. One can receive data not only from self-study of his case usage patterns, but also from asking others to review his communication efforts and give him their reactions. Figure 5.3 illustrates the essential cases used by each of the speakers in the sample dialogue.

5.2.3 Use of Peripheral Cases with each Verb Type Much of what we have said about essential cases also applies to the use of peripheral cases. How a speaker uses these cases also provides participants and observers with data for inferences. Depending on the verb type and the lexical entry of that verb type, peripheral cases supply the observer with various kinds of data.

For example, if a speaker is discussing an event it seems that the more peripheral cases he uses the more details he is providing. Note that the first sentence of the sample dialogue contains three peripheral cases. However, if the speaker is promising to perform a service or to act in some manner, a proliferation of peripheral cases indicates a series of qualifications about the commitment. For example, take the promise:

> "I will serve the organization in my spare time in this area without traveling and with other workers."

I will serve the organization	A O
in my spare time	T
in this area	L
without traveling	M
with other workers	C

The qualifications or details that a speaker evidences through use of peripheral cases provides an important source of data. Figure 5.4 illustrates the peripheral cases used in the sample dialogue.

5.2.4 Comparisons of Speakers' Use of Potential Information Units In Figures 5.2, 5.3, and 5.4 we have illustrated both speakers' usage of verb types, essential cases and peripheral cases. Speakers can be compared on each of these, all of these, or on selected ones across

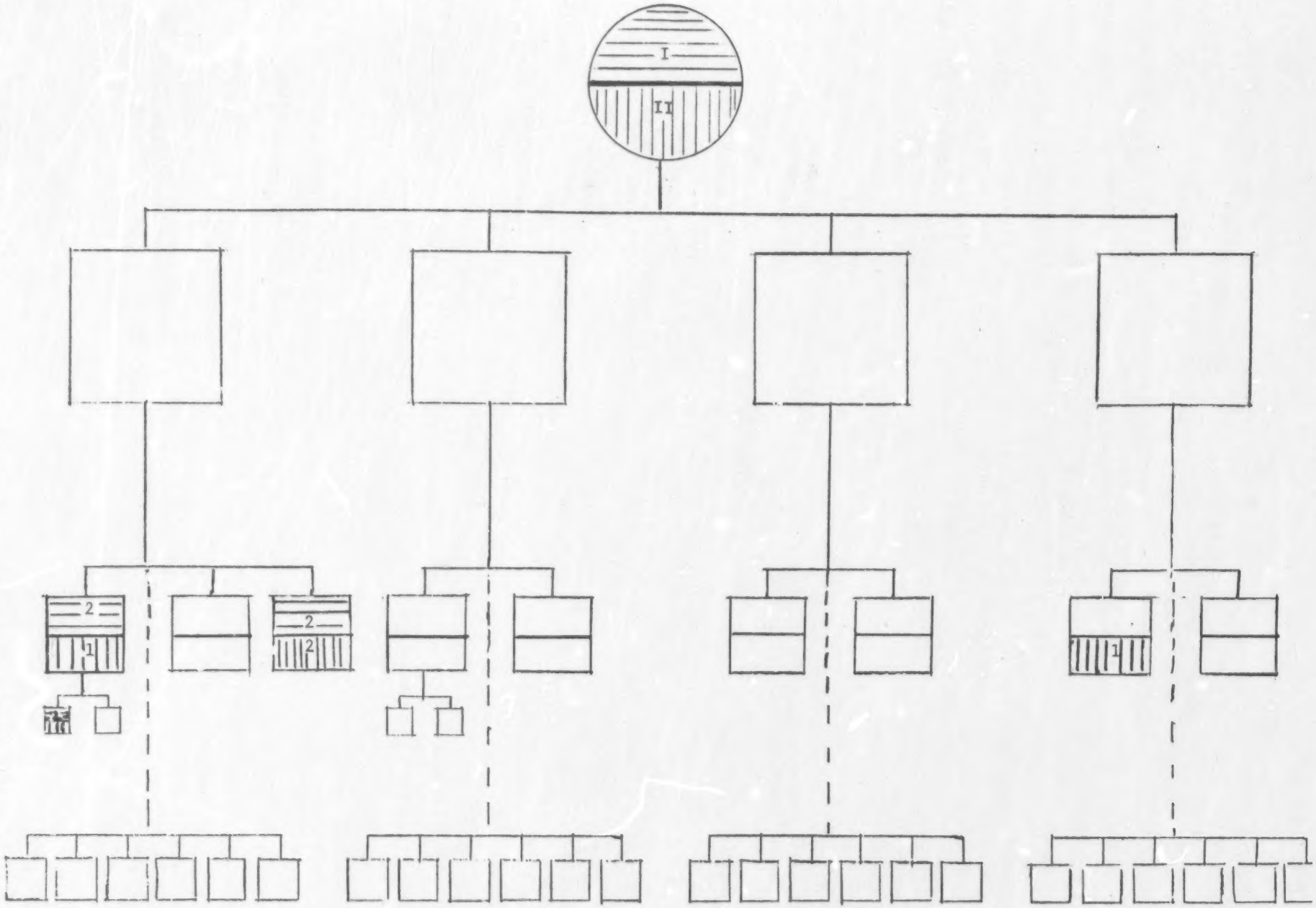


Figure 5.3 Essential Case Use in Sample Dialogue

(Arabic unmerals indicate the number of times the speaker used the essential case indicated.)

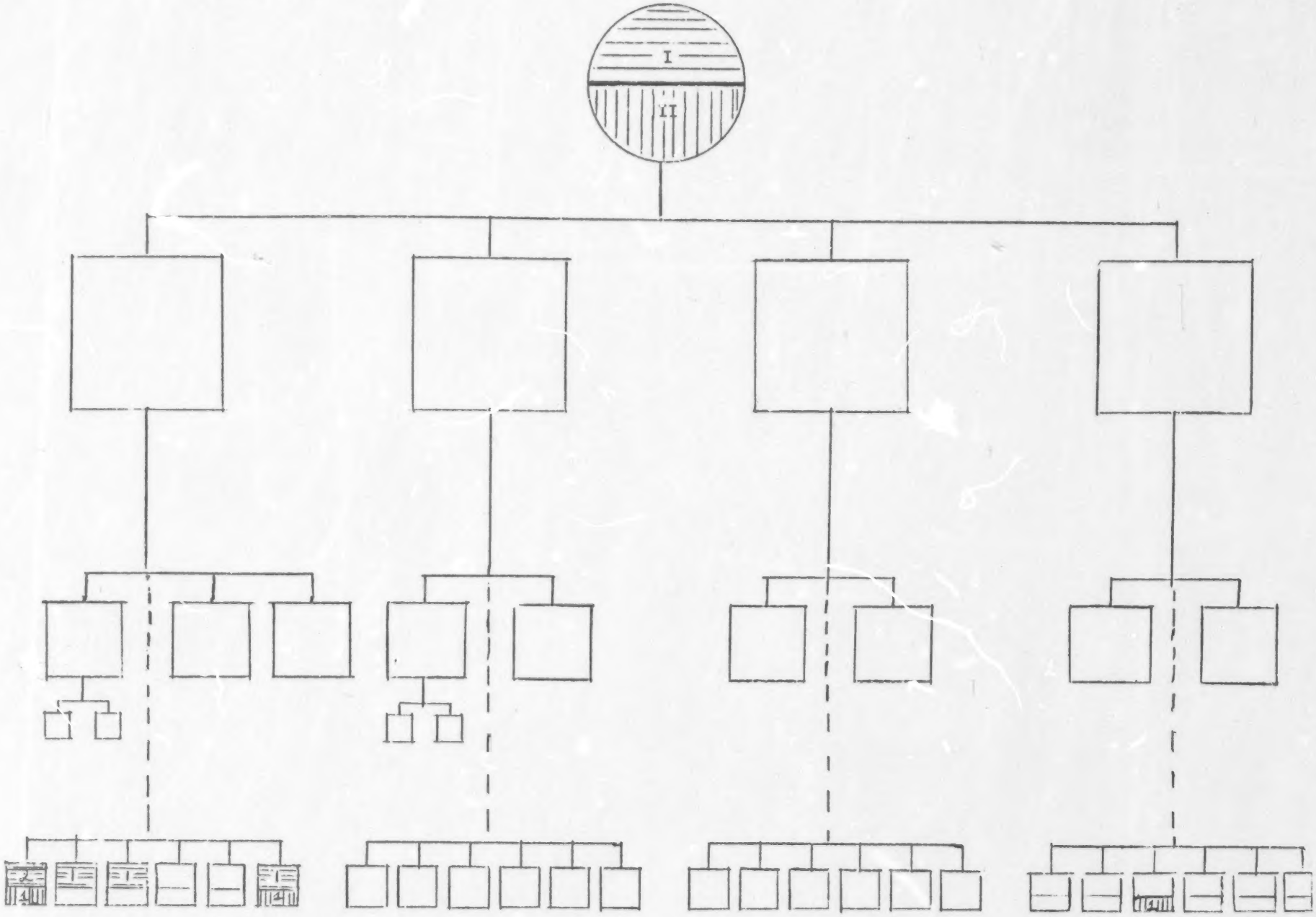


Figure 5.4 Peripheral Case Use in Sample Dialogue

(Arabic numerals indicate the number of times the speaker used the peripheral case indicated.)

specific topics. A case role profile can be prepared for each speaker and these profiles can be compared not only with the areas mentioned above but also with outside measures of the dialogue. After interpretation, a speaker's profile could be used as a predictive instrument, as well as for feedback to the individual himself and to other participants in the dialogue. Figure 5.5 indicates the speakers' profiles from the sample dialogue.

5.2.5 Lexicon and Use of Potential Information Units Sometimes it may be of use to observers to sort the dialogue to determine if certain lexical entries are more likely to produce a particular case role pattern. The observer may not be particularly interested in the dialogue as a whole, but rather how the participants handle certain topics or ideas.

5.2.6 Sequence and Use of Potential Information Units As dialogue progresses the data generated may also change. What each speaker says is determined in part by what others say (see Figure 2.1). It may be important therefore to sample potential information units at different times in the dialogue, and to compare the results. The differences among comparisons will depend upon many factors including how well the participants believe the discussion to be proceeding, what, if any, goals the participants have, whether a topical agenda is being used, and so forth.

5.2.7 Summary Users of CALAS must determine for themselves what potential information units from the dialogue analysis are most useful for their own purposes. For those who use CALAS in long term research programs, comparison of potential information units with outside measures may prove useful in initial stages. The units themselves can be used as feedback to participants concerning their own performance as well as data from which observers can make inferences. Figure 5.5 presents speakers' profiles from the sample dialogue.

6. Future Plans

Future work with CALAS is being planned in two major areas: (1) further refinements in the system; (2) empirical use of the system in dialogue analysis. Each of these areas is discussed in turn.

6.1 Further Refinement for CALAS

In addition to minor refinements in coding and other measures to improve CALAS' efficiency three major refinements are planned: (1) a reflexive case; (2) treatment of process and intransitive agent verbs, and (3) case role assignment clauses. There are occasions when a speaker or writer uses a reflexive construction in his utterances to indicate that the instigator and recipient of the action are the same person. For example, note the following:

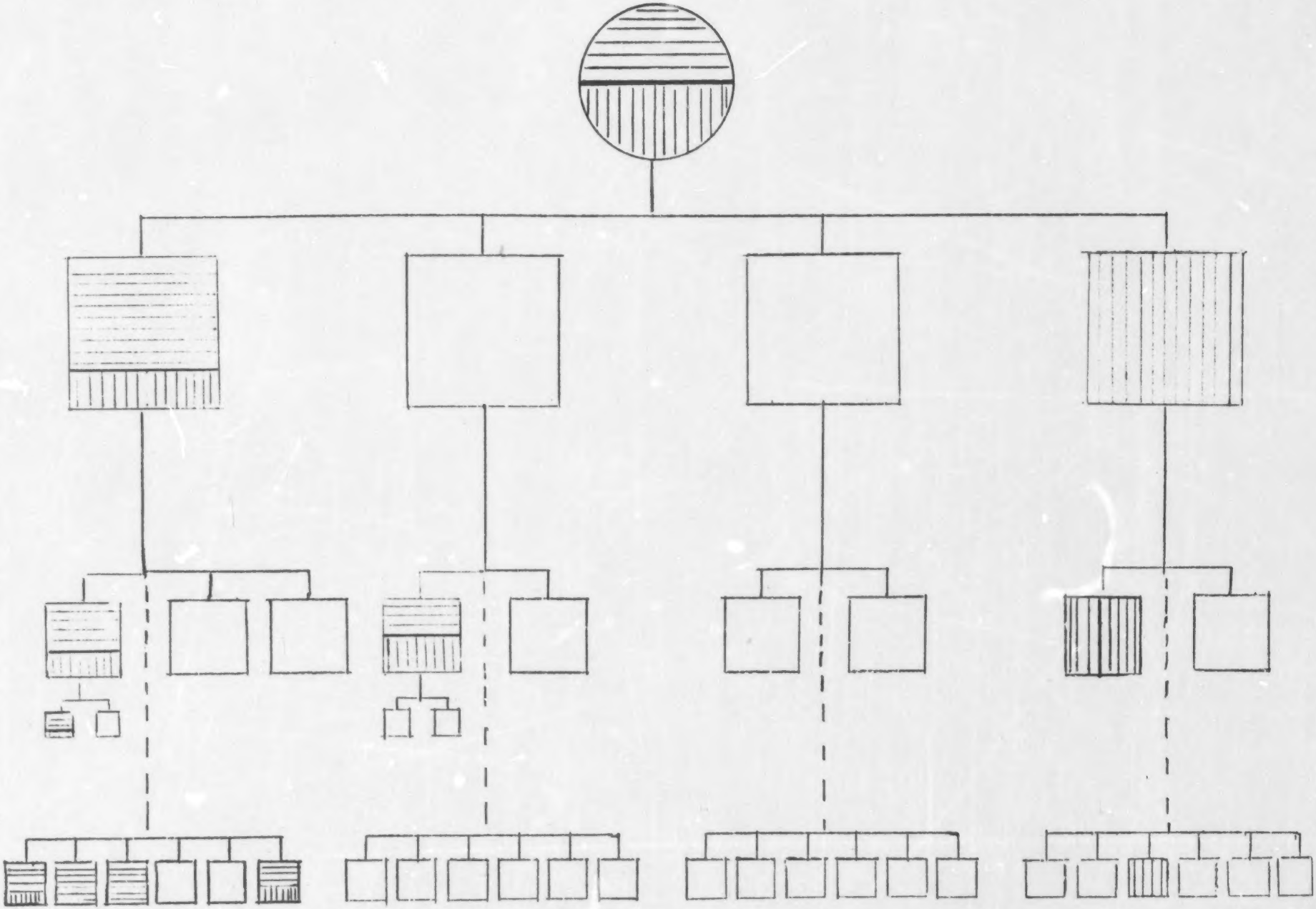


Figure 5.5 Speakers' Profiles from Sample Dialogue

I injured myself.

He laughed himself silly.

We worked ourselves too hard.

A new case role, Reflexive, has been developed to describe such examples. A case frame [R R] algorithm will be implemented in PL/I to note reflexive constructions. The R-R case frame is much like the O-O in that each indicates that the noun phrase before and the noun phrase after the verb refer to the same person or object.

As noted earlier the verb classifications made by CALAS do not include process verbs. Process verbs are those which indicate that the noun phrase which fills the subject slot is the receiver of the action or process which is indicated by the verb. For example, note the following:

A tree grows in Brooklyn.	[O - L]
The cake baked in the oven.	[O - L]
The plot developed slowly.	[O - M]

In each instance the underlined noun phrase receives an object case. There are no structural cues to indicate this case role assignment. Those same verbs when used transitively change the case role assignments.

Joe grows tomatoes.	[A - O]
Mother bakes cakes.	[A - O]
The chemicals developed the pictures.	[A - O]

By the same token all intransitive verbs are not process verbs

Mary ran down the street.	[A L]
The sun shown brightly.	[A M]
The painters finished yesterday.	[A T]

It is very difficult, therefore, to know what case roles to assign unless the lexical "meanings" of the verbs are known. And in some cases this is not enough. For instance in the sentence: "Mother baked", unless the context is known an observer does not know whether to infer that mother has made some cookies or was made uncomfortable by excessive heat.

We are now devising an algorithm to correct some of these difficulties in automatic verb classification and role assignment. This algorithm is based on the assumption that case roles are shaded by their environments. The reader has probably already noted that the object case can appear in several environments, and while it is an object in each instance we interpret it in terms of its verb type and other essential cases, if any, around it. Intransitivity then is an environment. It is an environment

that can be deduced from structure whatever the lexical "meaning" of the verb. We, therefore, propose another new case category to describe the subject slot in a clause where the verb is intransitive and not experiencer, stative or benefactive. This slot denotes that what appears in it is a mixture of instigator-participant-receiver of the action described by the verb. The algorithm then combines all these intransitive verbs and assigns their subject slot one case role. The case role has not as yet been named.

While there is a distinction between what we have called process verbs and intransitive agent verbs, computers cannot make this distinction at their present level of sophistication. And we think these verbs are more similar to each other than different, and more different from the transitive verbs than similar to them and therefore merit special case role consideration. If a CALAS user wishes to separate the two verb types by hand, he can do so by having each instance of this case role assignment automatically sorted and then make whatever distinction he wishes.

We do not anticipate the designation of additional case roles other than the two just mentioned. With these two additions now in the planning stages, CALAS will designate six essential cases and six peripheral ones. These twelve cases provide potential information units from which inferences can be made. There are plans, however, to extend case role assignments to dependent clauses. These clauses include subordinated clauses, relative clauses, and partial clauses. Every dependent clause has a function, and that function is determined by its relationship to the central verb in the independent clause. If that function is nominal or adverbial a case role assignment is indicated for the dependent clause. The case role assignments for dependent clauses are based on the same algorithms as the case role assignment for phrases. The case roles have the same labels and definitions. When the algorithm for case role assignment of dependent clauses is implemented, each phrase within the dependent clause will retain its case role and additional case roles will be assigned to the entire dependent clause. Figure 6.1 illustrates this concept.

In addition to these major improvements, minor refinements will also be made in the algorithms and their implementation. More work will be done to display more fully certain aspects of the language analysis. For example, special attention will be paid to interrogative and imperative mood utterances. These utterances have the same case roles as those in the indicative mood, but they need to be displayed differently.

6.2 Empirical Use of CALAS

CALAS has now been developed to such a stage of refinement that it is ready for use in language analysis studies. Those interested in investigating literary style and fact-to-face interaction may find CALAS particularly helpful (25).

At the moment validation and cross validation studies are needed to demonstrate the usefulness of this analytic instrument. Its reliability has already been established. The first step in this validation process will be a series of empirical studies correlating the CALAS analysis with outside measures of language evaluation. One such outside measure may be

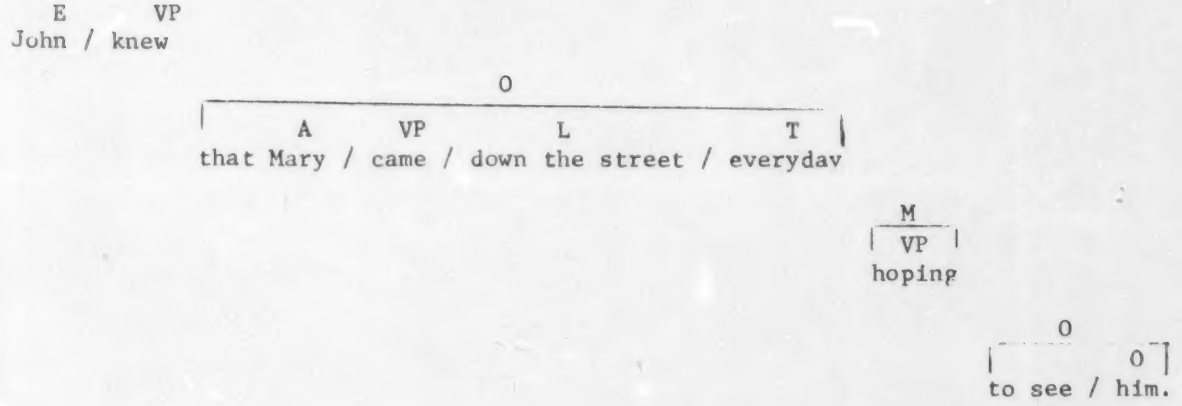


Figure 6.1 Case Assignment of Dependent Clauses and Phrases

the semantic differential (26). Another might be the participants' views of the dialogue. In text writing the author's stated intention may be compared with the analysis from CALAS.

After the initial series of exploratory studies, a series of more detailed ones will follow. In this series specific hypotheses derived from the results of the first series will be tested. These hypotheses will focus on determining which specific signals, isolated by CALAS, have predictive value. A later series of studies will ask subjects to manipulate these signals deliberately to determine if based on these manipulations outcomes can be predicted.

As discussed throughout the report, CALAS has many potential uses for instruction in communication and may be so used before these empirical validation studies are begun. CALAS provides a rapid reliable method for parsing language strings into their functions or roles. It is our belief that an understanding of these roles and their relationships can make attempts to interpret dialogue more accurate, useful and heuristic.

REFERENCES

(1) C. J. Fillmore, "The Case for Case", in E. Bach and R. Harms (eds.), Universals in Linguistic Theory., Holt, Rinehart and Winston, New York, 1968, 1-88.

(2) W. L. Chafe, Meaning and the Structure of Language., University of Chicago Press, Chicago, Ill., 1970.

(3) W. A. Cook, S. J., "Improvements in Case Grammar 1970", Language and Linguistics Working Papers, No. 2., Georgetwon University Press, Washington, D.C. 1971, 10-23.

(4) H. Garfinkel, Studies in Ethnomethodology., Prentice-Hall, Englewood Cliffs, 1967.

(5) M. C. Yovits and R. L. Ernst, "Generalized Information Systems: Consequences for Information Transfer", in H. B. Pepinsky (ed.), People and Information. Perganon Press, Elmsford, New York, 1970, 1-31.

(6) W. A. Cook, S. J., Introduction to Tagmemic Analysis., Holt, Rinehart and Winston, Inc., New York, 1969.

(7) B. C. Landry and J. E. Rush, "Toward a Theory of Indexing - II", Journal of the American Society for Information Science., 21 (5), 1970, 358-367.

(8) B. C. Landry, A Theory of Indexing: Indexing Theory as a Model for Information Storage and Retrieval., Computer and Information Science Research Center, CISRC-TR-71-13, The Ohio State University, 1971.

(9) C. J. Fillmore, "A Proposal Concerning English Prepositions", Georgetown University Monograph Series on Languages and Linguistics, No. 19. Georgetown University Press, Washington, D.C., 19-33, 1966.

(10) C. J. Fillmore, "Toward a Modern Theory of Case", In: D. A. Reibel and S. A. Schane (eds.), Modern Studies in English. Prentice-Hall, Inc. 361-375, 1969.

(11) C. J. Fillmore, "Some Problems for Case Grammar", Georgetown University Monograph Series on Languages and Linguistics, No. 22., Georgetown University Press, Washington, D.C., 35-56, 1971.

(12) W. A. Cook, S. J., "Case Grammar: From Roles to Rules", Languages and Linguistics Working Papers, No. 1., Georgetown University Press, Washington, D.C., 14-29, 1970.

(13) W. A. Cook, S. J., "A Set of Postulates for Case Grammar Analysis", Languages and Linguistics Working Papaer, No. 4., Georgetwon University Press, Washington, D.C., 35-49, 1972.

(14) W. A. Cook, S. J., "A Case Grammar Matrix", Languages and Linguistics Working Papers, No. 6., Georgetwon University Press, Washington, D.C., 15-47, 1972.

(15) D. M. Lambert, "The Semantic Syntax of Metaphor: A Case Grammar Analysis", Unpublished Ph.D. Dissertation, University of Michigan, 1969.

(16) A. M. McCoy, "A Case Grammar Classification of Spanish Verbs", Unpublished Ph.D. Dissertation, University of Michigan, 1969.

(17) D. L. F. Nilsen, "Toward a Semantic Specification of Deep Case", Unpublished Ph.D. Dissertation, University of Michigan, 1971.

(18) J. M. Anderson, The Grammar of Case: Towards a Localistic Theory., Cambridge University Press, Cambridge, 1971.

(19) F. M. Aid, Semantic Structures in Spanish: A Proposal for Instructional Materials., Georgetwon University Press, Washington, D.C., 1973.

(20) C. E. Young, "Grammatical Assignment Based on Function Words". Unpublished Master's Thesis, The Ohio State University, 1970.

(21) P. J. Stone, D. C. Dunphy, M. S. Smith, D. M. Ogilvie, The General Inquirer: A Computer Approach to Content Analysis., The M.I.T. Press, Cambridge, Massachusetts, 1966.

(22) H. Kucera and W. R. Francis, Computational Analysis of Present Day American English., Brown University, Providence, R. I., 1967.

(23) W. A. Cook, S. J., Private Communication 1971

(24) E. Hemingway, The Old Man and the Sea, Charles Scribner's Sons, New York, New York, 1952.

(25) H. B. Pepinsky, "A Metalanguage for Systematic Research on Human Communication Via Natural Language", Journal of the American Society for Information Science, (in press).

(26) C. E. Osgood, G. J. Suci, P. H. Tannenbaum, The Measurement of Meaning., University of Illinois Press, Urbana, Illinois, 1957.

COMPUTER &
INFORMATION
SCIENCE
RESEARCH CENTER